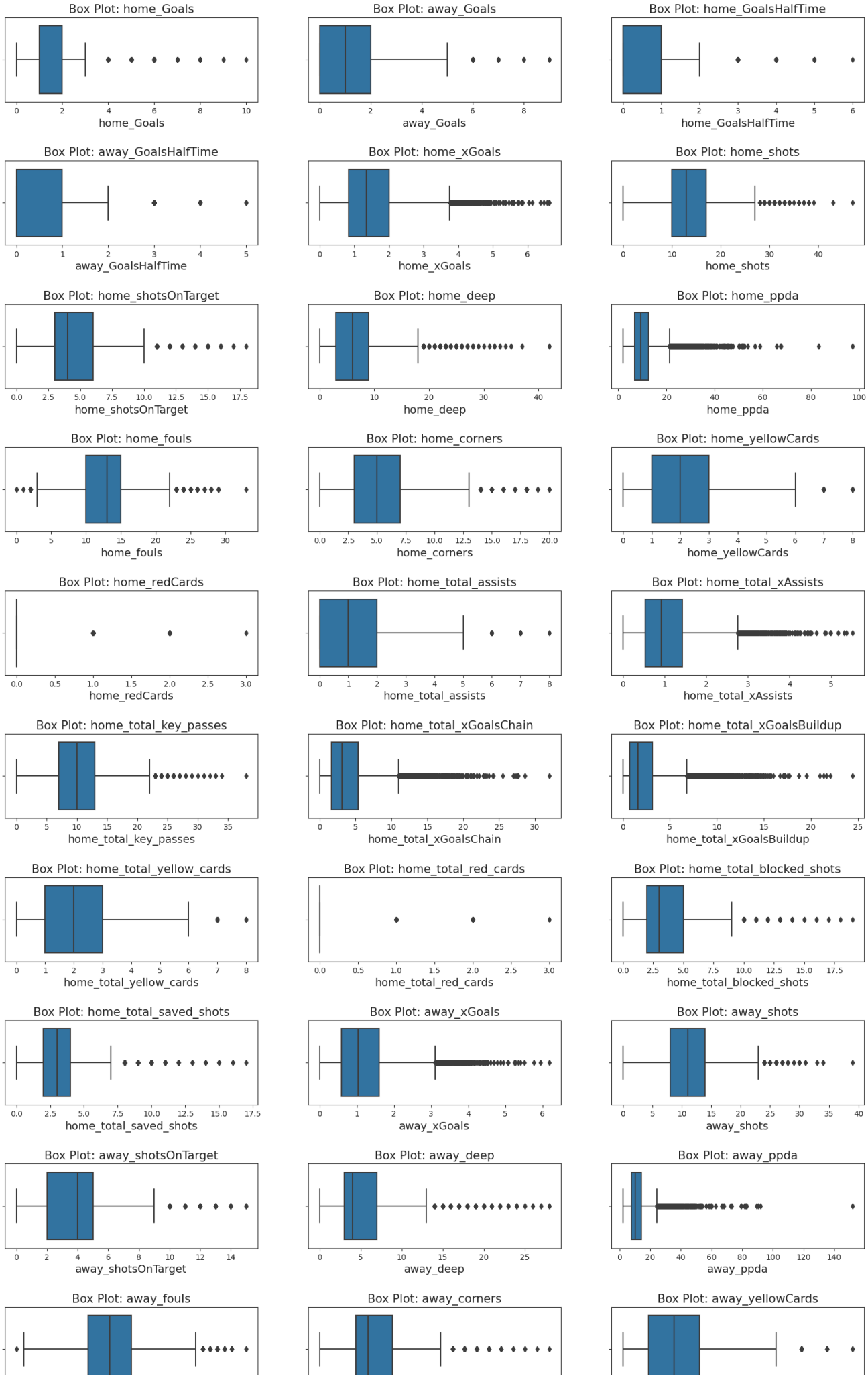
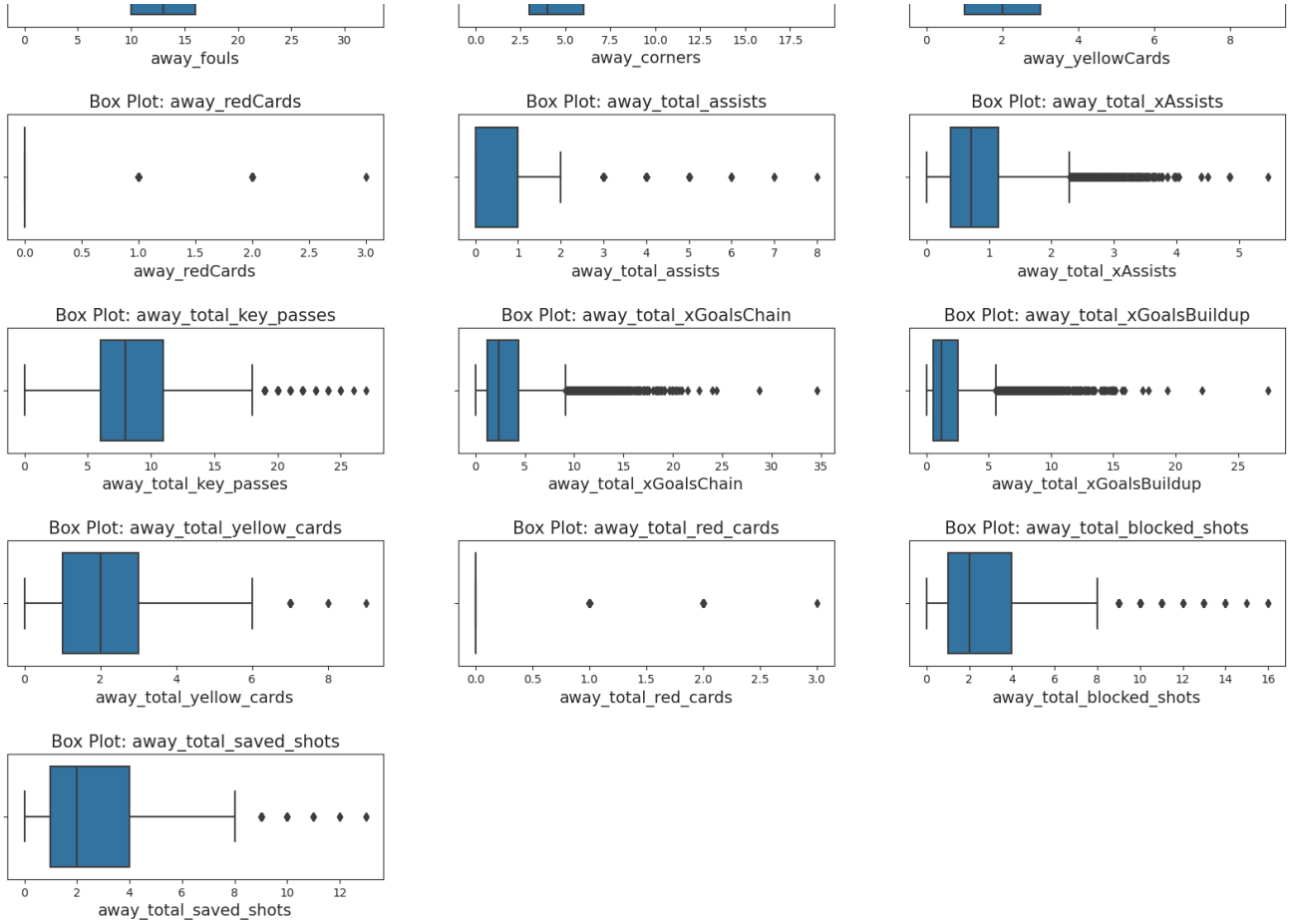
loading and importing

	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
0	81	1	2015	2015-08-08 15:45:00	89	82	1	0
1	82	1	2015	2015-08-08 18:00:00	73	71	0	1
2	83	1	2015	2015-08-08 18:00:00	72	90	2	2
3	84	1	2015	2015-08-08 18:00:00	75	77	4	2
4	85	1	2015	2015-08-08 18:00:00	79	78	1	3
...	...	...	...	...	...	...	...	...
12675	16131	5	2020	2021-05-23 19:00:00	168	166	1	2
12676	16132	5	2020	2021-05-23 19:00:00	177	176	1	2
12677	16133	5	2020	2021-05-23 19:00:00	163	235	2	0
12678	16134	5	2020	2021-05-23 19:00:00	175	181	0	1
12679	16135	5	2020	2021-05-23 19:00:00	225	179	1	1

12680 rows × 51 columns

Outliers


Box Plot: home_Goals
home_Goals
Box Plot: away_Goals
away_Goals
Box Plot: home_GoalsHalfTime
home_GoalsHalfTime
Box Plot: away_GoalsHalfTime
away_GoalsHalfTime
Box Plot: home_xGoals
home_xGoals
Box Plot: home_shots
home_shots
Box Plot: home_shotsOnTarget
home_shotsOnTarget
Box Plot: home_deep
home_deep
Box Plot: home_ppda
home_ppda
Box Plot: home_fouls
home_fouls
Box Plot: home_corners
home_corners
Box Plot: home_yellowCards
home_yellowCards
Box Plot: home_redCards
home_redCards
Box Plot: home_total_assists
home_total_assists
Box Plot: home_total_xAssists
home_total_xAssists
Box Plot: home_total_key_passes
home_total_key_passes
Box Plot: home_total_xGoalsChain
home_total_xGoalsChain
Box Plot: home_total_xGoalsBuildup
home_total_xGoalsBuildup
Box Plot: home_total_yellow_cards
home_total_yellow_cards
Box Plot: home_total_red_cards
home_total_red_cards
Box Plot: home_total_blocked_shots
home_total_blocked_shots
Box Plot: home_total_saved_shots
home_total_saved_shots
Box Plot: away_xGoals
away_xGoals
Box Plot: away_shots
away_shots
Box Plot: away_shotsOnTarget
away_shotsOnTarget
Box Plot: away_deep
away_deep
Box Plot: away_ppda
away_ppda
Box Plot: away_fouls
Box Plot: away_corners
Box Plot: away_yellowCards



the missing value in homeYellowCards was dealt in the EDA analysis when creating the categorical features.

Creating Nulls dataframe and matrix

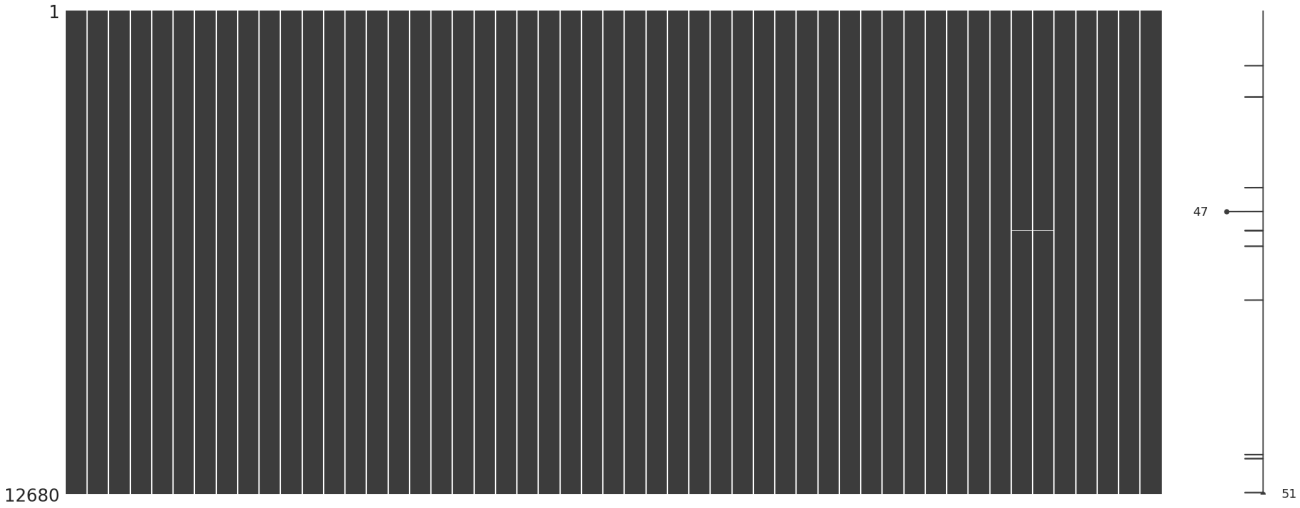
	home_total_blocked_shots	home_total_saved_shots	away_total_blocked_shots	away_total_
0	4.0	1.0	3.0	
1	2.0	2.0	2.0	
2	2.0	3.0	3.0	
3	4.0	4.0	2.0	
4	3.0	4.0	2.0	
...	...	...	...	
12675	6.0	4.0	1.0	
12676	3.0	2.0	4.0	
12677	2.0	4.0	0.0	
12678	5.0	5.0	1.0	
12679	0.0	1.0	2.0	

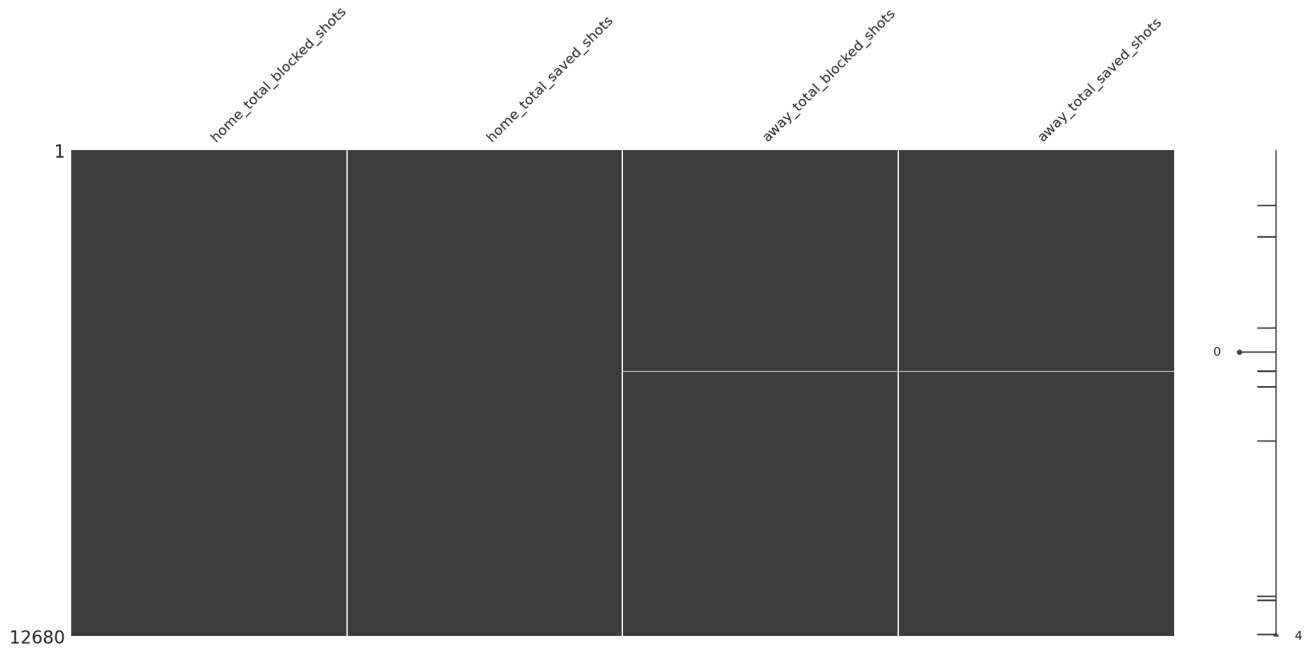
12680 rows × 4 columns

```
home_total_blocked_shots  home_total_saved_shots  away_total_blocked_shots  away_t
otal_saved_shots
2.0                       2.0                     2.0                       3.0
31
                          3.0                     2.0                       3.0
28
3.0                       4.0                     2.0                       2.0
28
2.0                       2.0                     3.0                       3.0
27
                                                  2.0                       1.0
26


..
3.0                       9.0                     6.0                       4.0
1
                          10.0                    1.0                       0.0
1
                                                                            1.0
1
                                                                            3.0
1
NaN                       NaN                     NaN                       NaN
1
Length: 4079, dtype: int64

<Axes: >
<Axes: >
```





Dataframe containing missing value counts and their frequency:

	Missing Values	% of Total Values
away_total_blocked_shots	8	0.1
away_total_saved_shots	8	0.1
home_total_blocked_shots	3	0.0
home_total_saved_shots	3	0.0

	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
5270	6018	5	2014	2015-01-18 20:00:00	164	169	2	1
7592	9486	1	2018	2019-03-02 15:00:00	73	88	0	1
11751	15207	3	2020	2020-11-21 14:30:00	262	119	1	2

3 rows × 51 columns

	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
5270	6018	5	2014	2015-01-18 20:00:00	164	169	2	1
7592	9486	1	2018	2019-03-02 15:00:00	73	88	0	1
11751	15207	3	2020	2020-11-21 14:30:00	262	119	1	2

3 rows × 51 columns

	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
1447	1528	4	2015	2015-11-29 23:30:00	138	146	1	0
2267	2348	5	2016	2016-10-23 22:45:00	161	164	0	0
4644	5392	3	2014	2014-10-18 14:30:00	117	123	6	0
5270	6018	5	2014	2015-01-18 20:00:00	164	169	2	1
5771	7413	1	2017	2018-03-10 15:00:00	219	84	0	0
6179	7821	2	2017	2018-04-17 18:45:00	106	116	4	0
11644	15100	4	2020	2021-04-18 14:15:00	143	156	5	0
12642	16098	5	2020	2021-05-01 19:00:00	160	170	2	0

8 rows × 51 columns

	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
1447	1528	4	2015	2015-11-29 23:30:00	138	146	1	0
2267	2348	5	2016	2016-10-23 22:45:00	161	164	0	0
4644	5392	3	2014	2014-10-18 14:30:00	117	123	6	0
5270	6018	5	2014	2015-01-18 20:00:00	164	169	2	1
5771	7413	1	2017	2018-03-10 15:00:00	219	84	0	0
6179	7821	2	2017	2018-04-17 18:45:00	106	116	4	0
11644	15100	4	2020	2021-04-18 14:15:00	143	156	5	0
12642	16098	5	2020	2021-05-01 19:00:00	160	170	2	0

8 rows × 51 columns

Creating a dataframe with each of the missing values as 1, while non missing values are 0:

	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
count	12680.0	12680.0	12680.0	12680.0	12680.0	12680.0	12680.0	12680.0
mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
std	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
50%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
75%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
max	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

8 rows × 51 columns

outliers_df

The data is skewed and not normaly distributed, we'll use the IQR method to identify outliers to set up a boundary before Q1 and after Q3. Any values that fall outside of

this boundary are considered outliers:

	Outlier count	Percent
away_redCards	1396.0	11.009464
away_total_red_cards	1382.0	10.899054
home_redCards	1078.0	8.501577
home_total_red_cards	1064.0	8.391167
home_Goals	981.0	7.736593
away_total_assists	790.0	6.230284
away_total_xGoalsBuildup	718.0	5.662461
home_total_xGoalsBuildup	664.0	5.236593
away_ppda	606.0	4.779180
home_ppda	540.0	4.258675
away_total_xGoalsChain	538.0	4.242902
home_total_saved_shots	532.0	4.195584
home_total_xGoalsChain	523.0	4.124606
away_deep	423.0	3.335962
home_GoalsHalfTime	403.0	3.178233
away_total_xAssists	389.0	3.067823
home_shotsOnTarget	351.0	2.768139
home_total_xAssists	346.0	2.728707
away_xGoals	312.0	2.460568
away_corners	283.0	2.231861
home_xGoals	276.0	2.176656
home_fouls	241.0	1.900631
away_shotsOnTarget	233.0	1.837539
home_total_blocked_shots	230.0	1.813880
home_deep	227.0	1.790221
away_GoalsHalfTime	225.0	1.774448
away_total_blocked_shots	187.0	1.474763
away_total_key_passes	169.0	1.332808
home_shots	167.0	1.317035
away_shots	161.0	1.269716
home_corners	143.0	1.127760
home_total_key_passes	123.0	0.970032
away_total_saved_shots	95.0	0.749211
away_fouls	81.0	0.638801
away_Goals	44.0	0.347003
away_yellowCards	43.0	0.339117

	Outlier count	Percent
home_yellowCards	37.0	0.291798
away_total_yellow_cards	26.0	0.205047
home_total_yellow_cards	20.0	0.157729
home_total_assists	20.0	0.157729

new_outliers_df

	Outlier count	Percent
away_redCards	1396.0	11.009464
away_total_red_cards	1382.0	10.899054
home_redCards	1078.0	8.501577
home_total_red_cards	1064.0	8.391167
home_Goals	981.0	7.736593
away_total_assists	790.0	6.230284
away_total_xGoalsBuildup	718.0	5.662461
home_total_xGoalsBuildup	664.0	5.236593
away_ppda	606.0	4.779180
home_ppda	540.0	4.258675
away_total_xGoalsChain	538.0	4.242902
home_total_saved_shots	532.0	4.195584
home_total_xGoalsChain	523.0	4.124606
away_deep	423.0	3.335962
home_GoalsHalfTime	403.0	3.178233
away_total_xAssists	389.0	3.067823
home_shotsOnTarget	351.0	2.768139
home_total_xAssists	346.0	2.728707
away_xGoals	312.0	2.460568
away_corners	283.0	2.231861
home_xGoals	276.0	2.176656
home_fouls	241.0	1.900631
away_shotsOnTarget	233.0	1.837539
home_total_blocked_shots	230.0	1.813880
home_deep	227.0	1.790221
away_GoalsHalfTime	225.0	1.774448
away_total_blocked_shots	187.0	1.474763
away_total_key_passes	169.0	1.332808
home_shots	167.0	1.317035
away_shots	161.0	1.269716
home_corners	143.0	1.127760
home_total_key_passes	123.0	0.970032
away_total_saved_shots	95.0	0.749211
away_fouls	81.0	0.638801
away_Goals	44.0	0.347003
away_yellowCards	43.0	0.339117

	Outlier count	Percent
home_yellowCards	37.0	0.291798
away_total_yellow_cards	26.0	0.205047
home_total_yellow_cards	20.0	0.157729
home_total_assists	20.0	0.157729

Labeling every outlier with 'Outlier' in order to seperate the outliers from nulls (temporarily, and then to 0/1):

	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
0	81	1	2015	2015-08-08 15:45:00	89	82	1	0
1	82	1	2015	2015-08-08 18:00:00	73	71	0	1
2	83	1	2015	2015-08-08 18:00:00	72	90	2	2
3	84	1	2015	2015-08-08 18:00:00	75	77	Outlier	2
4	85	1	2015	2015-08-08 18:00:00	79	78	1	3
...	...	...	...	...	...	...	...	...
12675	16131	5	2020	2021-05-23 19:00:00	168	166	1	2
12676	16132	5	2020	2021-05-23 19:00:00	177	176	1	2
12677	16133	5	2020	2021-05-23 19:00:00	163	235	2	0
12678	16134	5	2020	2021-05-23 19:00:00	175	181	0	1
12679	16135	5	2020	2021-05-23 19:00:00	225	179	1	1

12680 rows × 51 columns

df_outliers

Creating a dataframe with outliers as 1 and non-outliers as 0:

	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals	ho
0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	1	0	
4	0	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	
12675	0	0	0	0	0	0	0	0	
12676	0	0	0	0	0	0	0	0	
12677	0	0	0	0	0	0	0	0	
12678	0	0	0	0	0	0	0	0	
12679	0	0	0	0	0	0	0	0	

12680 rows × 51 columns

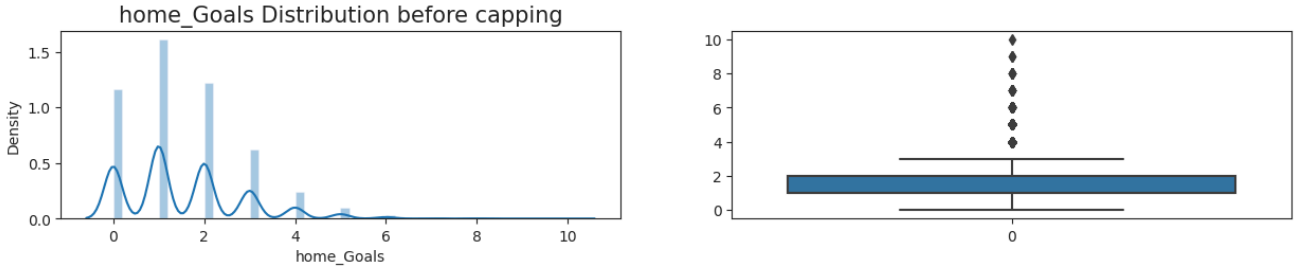
Now that we have a dataframe saving all the outliers we'll convert all the outliers to nulls, just so I can see differences in distribution with and without outliers and then decide which of the outliers to remove or not:

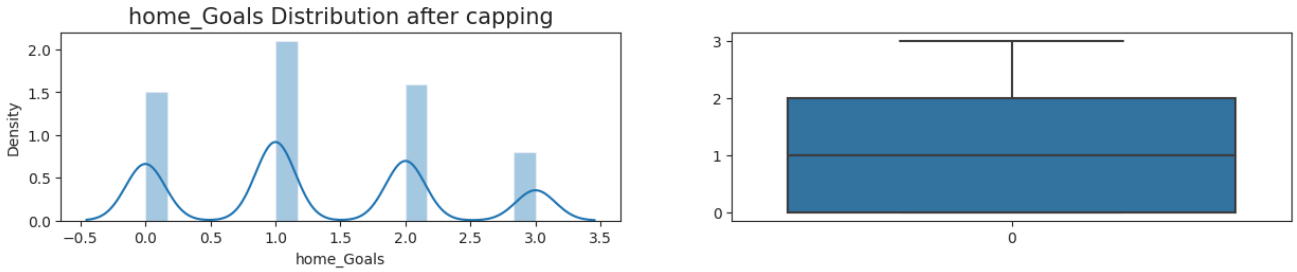
	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
0	81	1	2015	2015-08-08 15:45:00	89	82	1.0	0.0
1	82	1	2015	2015-08-08 18:00:00	73	71	0.0	1.0
2	83	1	2015	2015-08-08 18:00:00	72	90	2.0	2.0
3	84	1	2015	2015-08-08 18:00:00	75	77	NaN	2.0
4	85	1	2015	2015-08-08 18:00:00	79	78	1.0	3.0
...	...	...	...	...	...	...	...	...
12675	16131	5	2020	2021-05-23 19:00:00	168	166	1.0	2.0
12676	16132	5	2020	2021-05-23 19:00:00	177	176	1.0	2.0
12677	16133	5	2020	2021-05-23 19:00:00	163	235	2.0	0.0
12678	16134	5	2020	2021-05-23 19:00:00	175	181	0.0	1.0
12679	16135	5	2020	2021-05-23 19:00:00	225	179	1.0	1.0

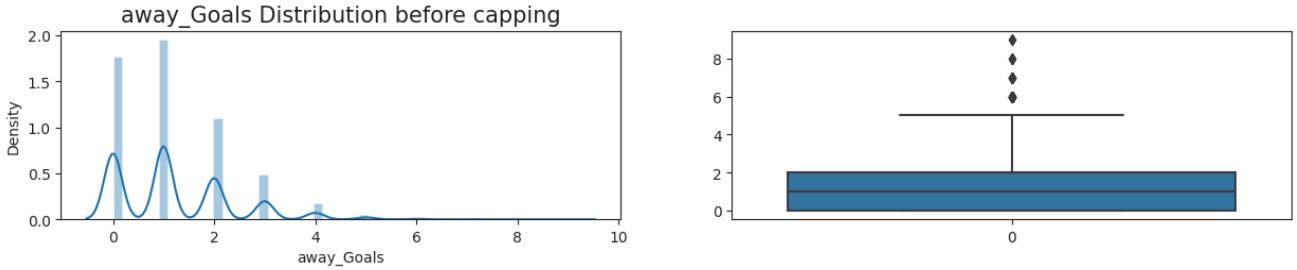
12680 rows × 51 columns

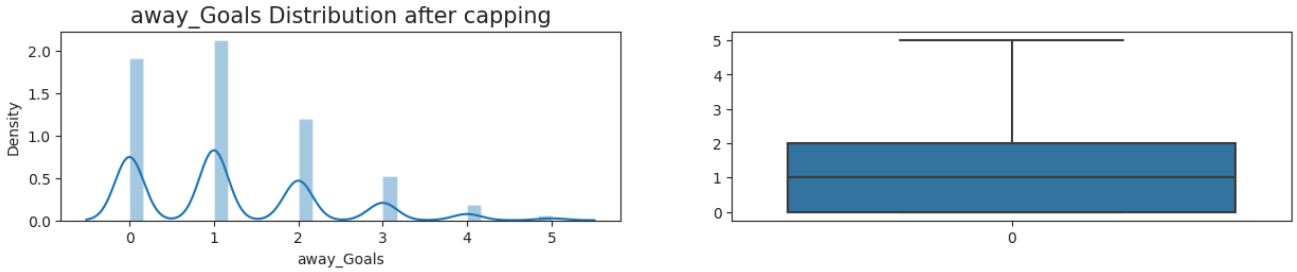
Saving the inter quartal outliers dataframe for future reference.

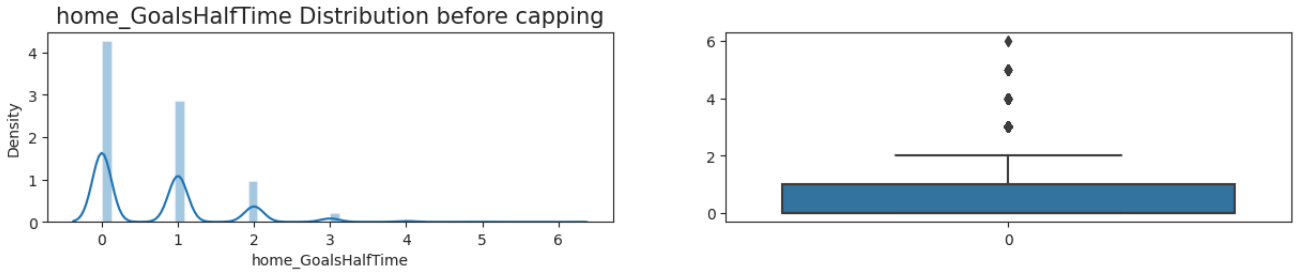
```
Index(['away_redCards', 'away_total_red_cards', 'home_redCards',
       'home_total_red_cards', 'home_Goals', 'away_total_assists',
       'away_total_xGoalsBuildup', 'home_total_xGoalsBuildup', 'away_ppda',
       'home_ppda', 'away_total_xGoalsChain', 'home_total_saved_shots',
       'home_total_xGoalsChain', 'away_deep', 'home_GoalsHalfTime',
       'away_total_xAssists', 'home_shotsOnTarget', 'home_total_xAssists',
       'away_xGoals', 'away_corners', 'home_xGoals', 'home_fouls',
       'away_shotsOnTarget', 'home_total_blocked_shots', 'home_deep',
       'away_GoalsHalfTime', 'away_total_blocked_shots',
       'away_total_key_passes', 'home_shots', 'away_shots', 'home_corners',
       'home_total_key_passes', 'away_total_saved_shots', 'away_fouls',
       'away_Goals', 'away_yellowCards', 'home_yellowCards',
       'away_total_yellow_cards', 'home_total_yellow_cards',
       'home_total_assists'],
      dtype='object')
```

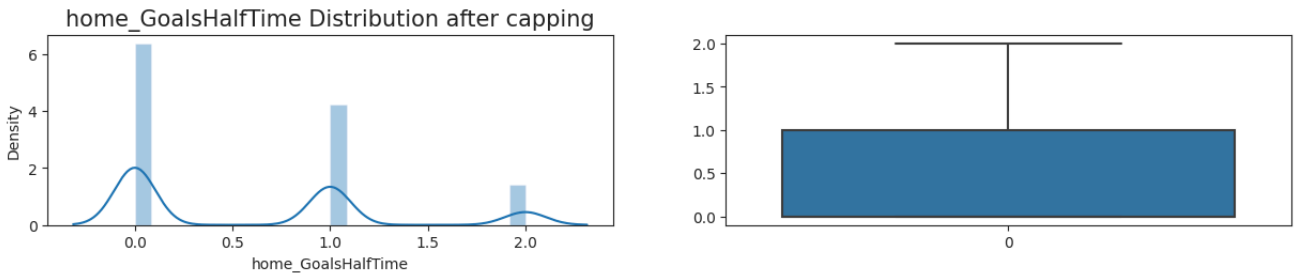

home_Goals Distribution before capping
Density
home_Goals

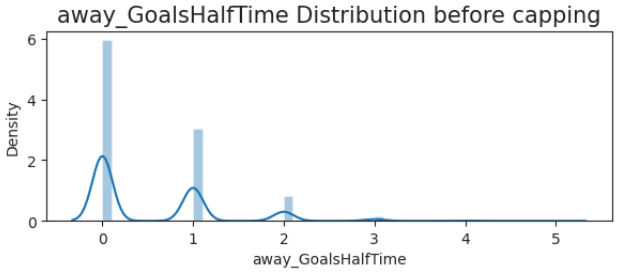

home_Goals Distribution after capping
Density
home_Goals

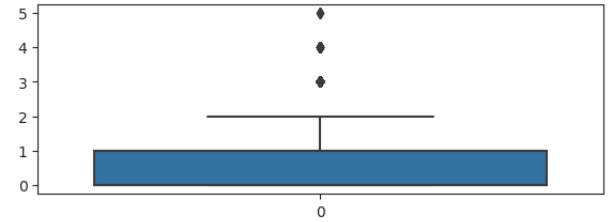

away_Goals Distribution before capping
Density
away_Goals

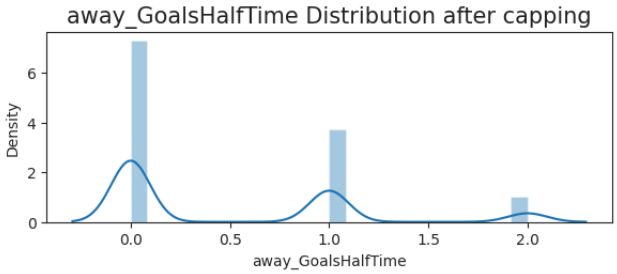

away_Goals Distribution after capping
Density
away_Goals

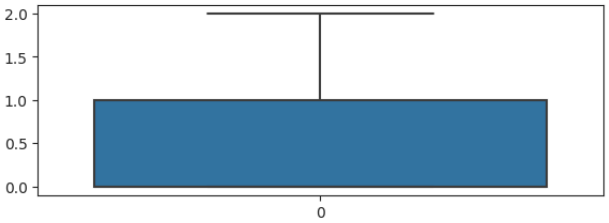

home_GoalsHalfTime Distribution before capping
Density
home_GoalsHalfTime

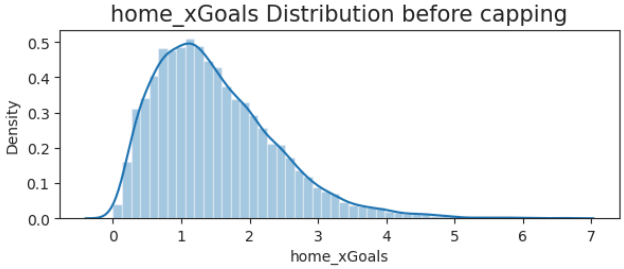

home_GoalsHalfTime Distribution after capping
Density
home_GoalsHalfTime

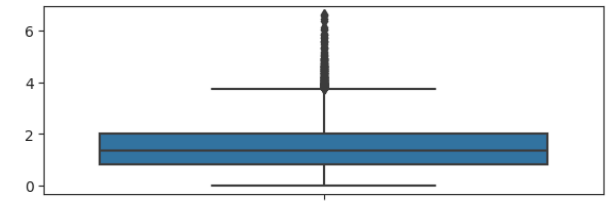

away_GoalsHalfTime Distribution before capping
Density
away_GoalsHalfTime

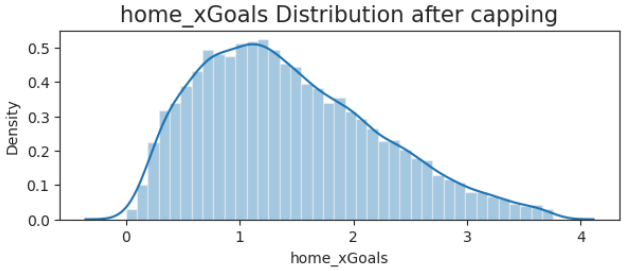


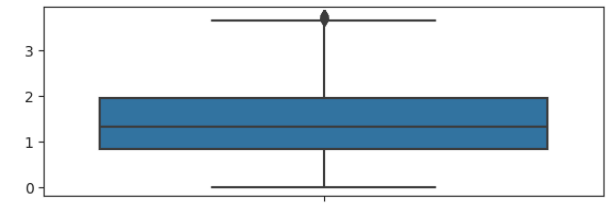

away_GoalsHalfTime Distribution after capping
Density
away_GoalsHalfTime

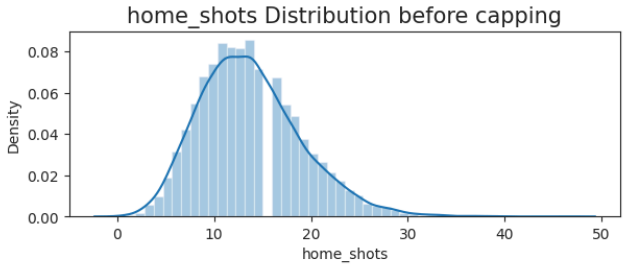


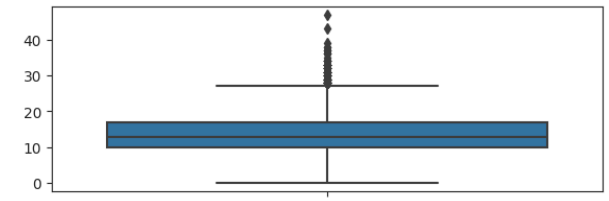

home_xGoals Distribution before capping
Density
home_xGoals

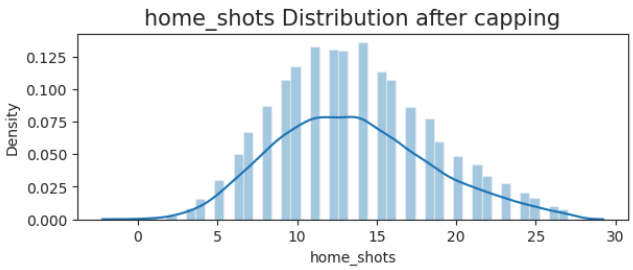


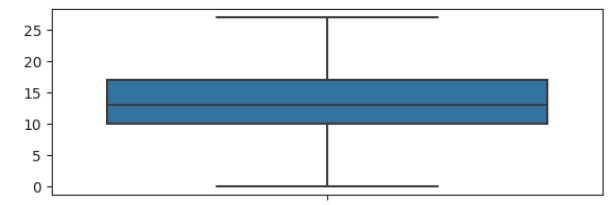

home_xGoals Distribution after capping
Density
home_xGoals

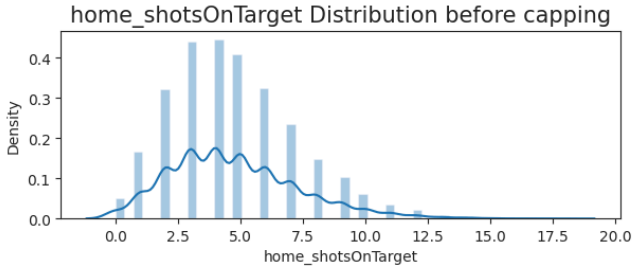


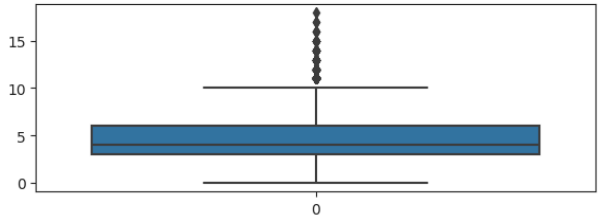

home_shots Distribution before capping
Density
home_shots

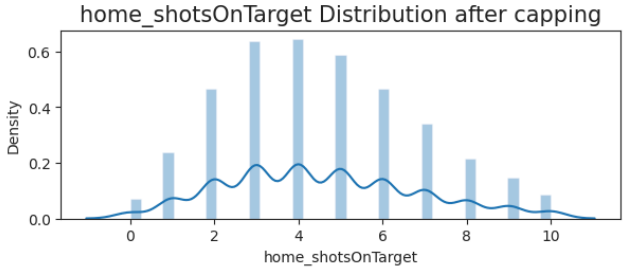


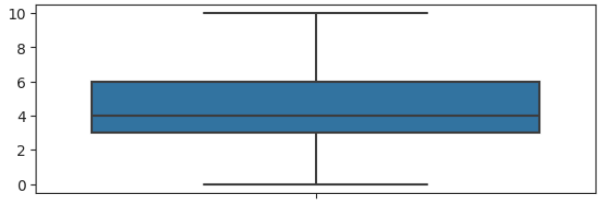

home_shots Distribution after capping
Density
home_shots

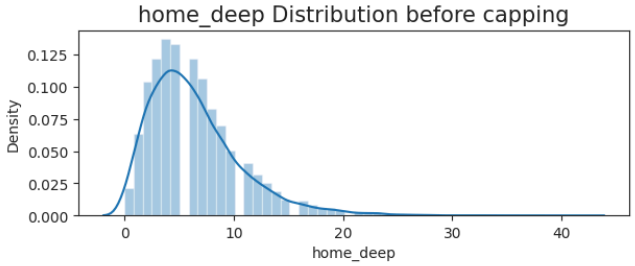


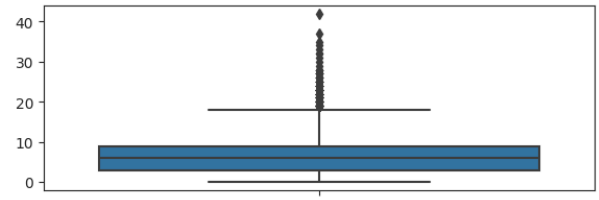

home_shotsOnTarget Distribution before capping
Density
home_shotsOnTarget

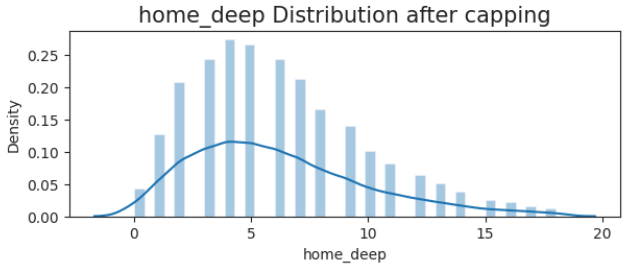


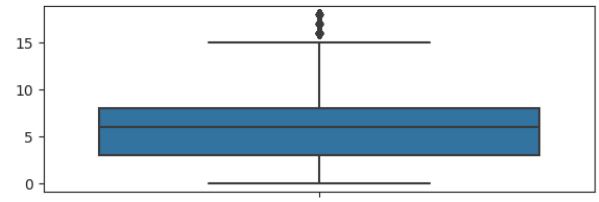

home_shotsOnTarget Distribution after capping
Density
home_shotsOnTarget

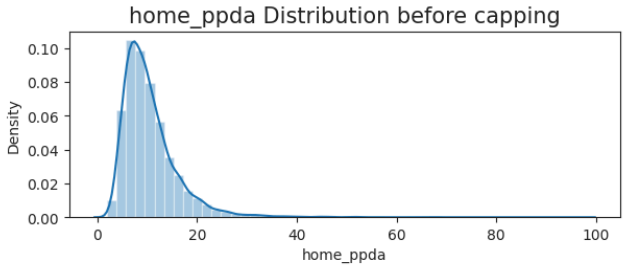


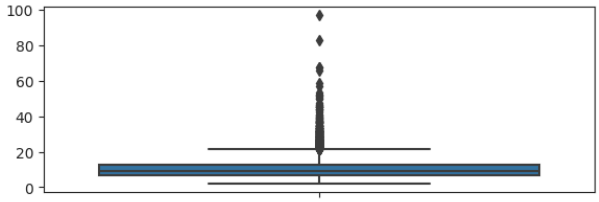

home_deep Distribution before capping
Density
home_deep

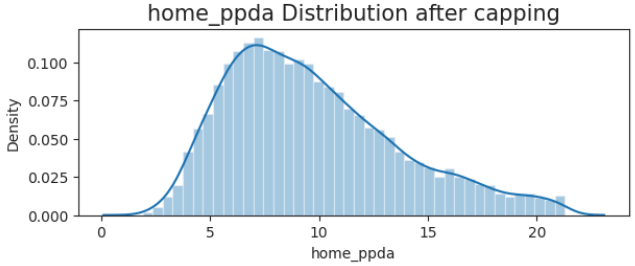


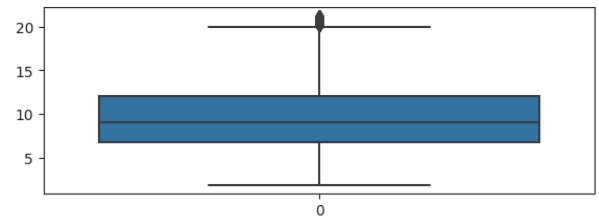

home_deep Distribution after capping
Density
home_deep

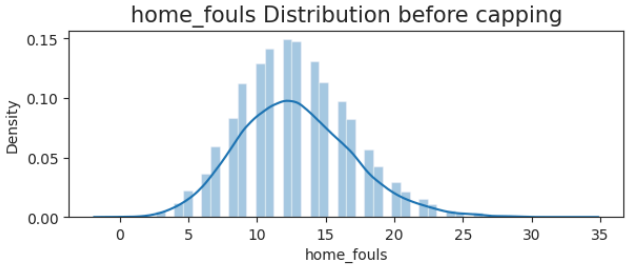


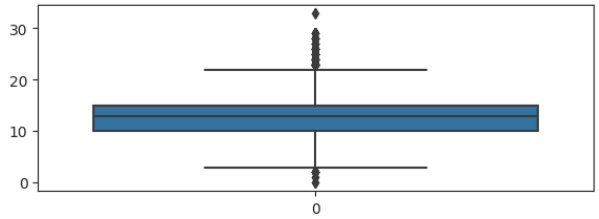

home_ppda Distribution before capping
Density
home_ppda

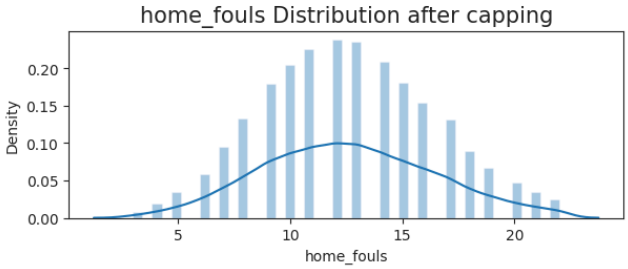


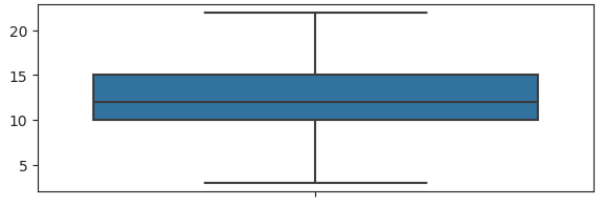

home_ppda Distribution after capping
Density
home_ppda

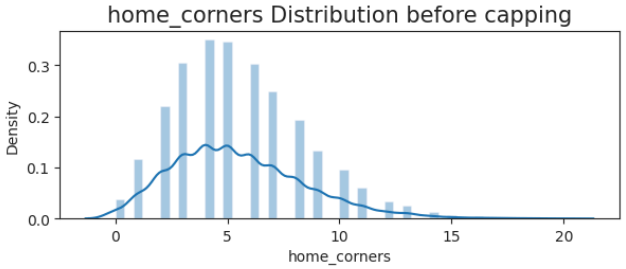


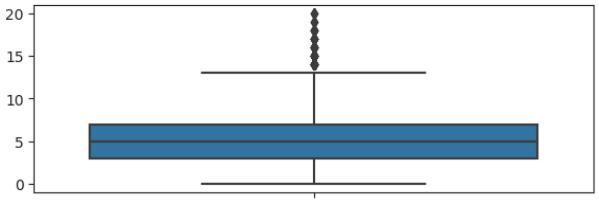

home_fouls Distribution before capping
Density
home_fouls

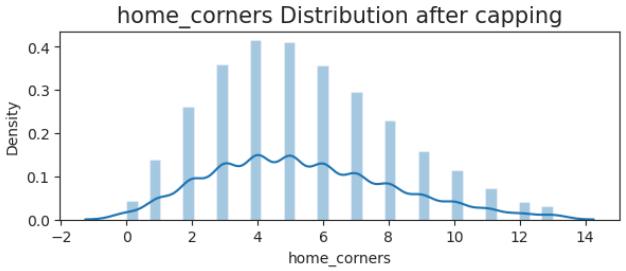


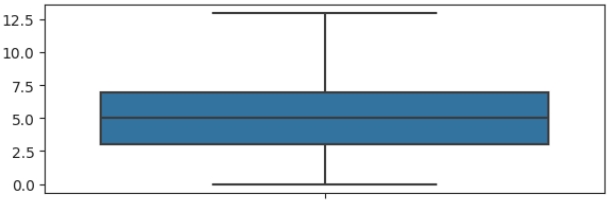

home_fouls Distribution after capping
Density
home_fouls

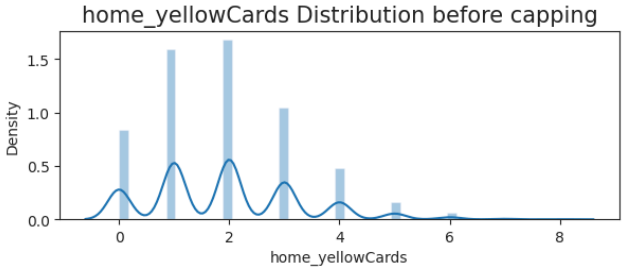


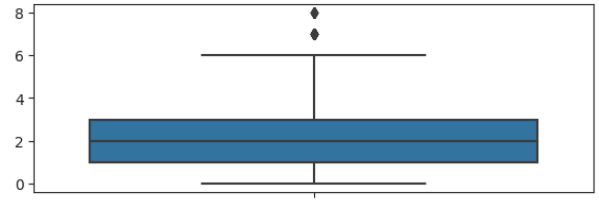

home_corners Distribution before capping
Density
home_corners

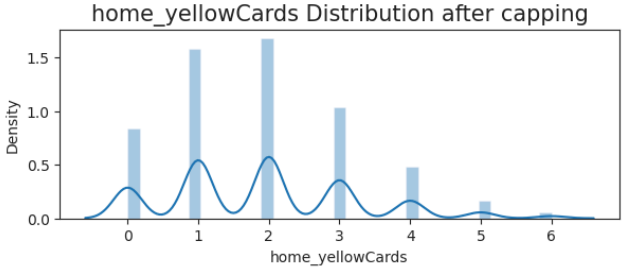


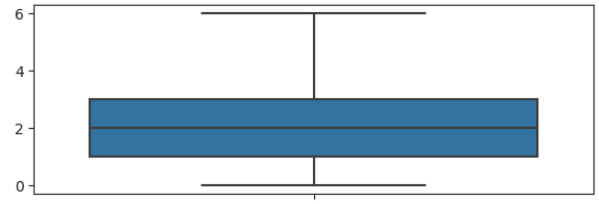

home_corners Distribution after capping
Density
home_corners

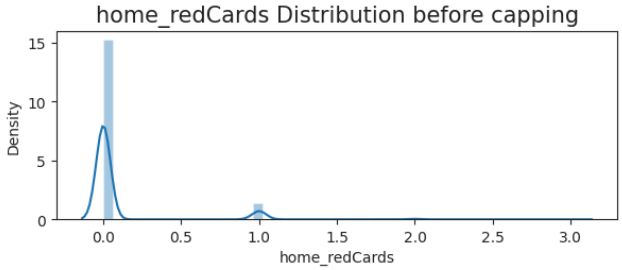


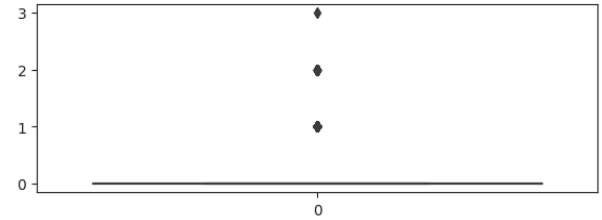

home_yellowCards Distribution before capping
Density
home_yellowCards

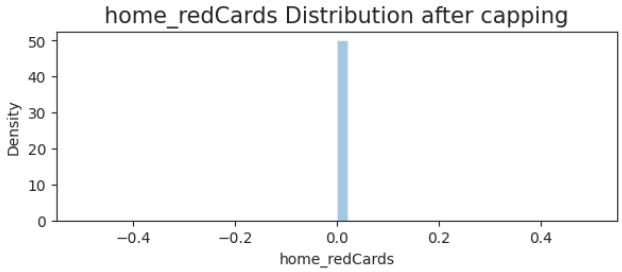


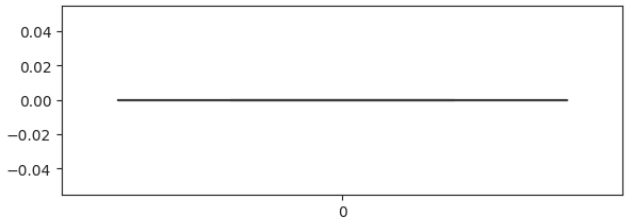

home_yellowCards Distribution after capping
Density
home_yellowCards

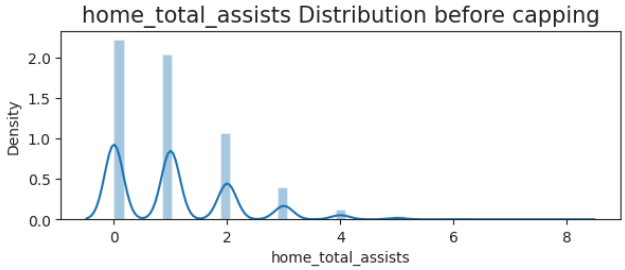


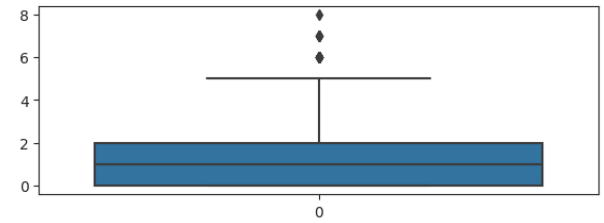

home_redCards Distribution before capping
Density
home_redCards

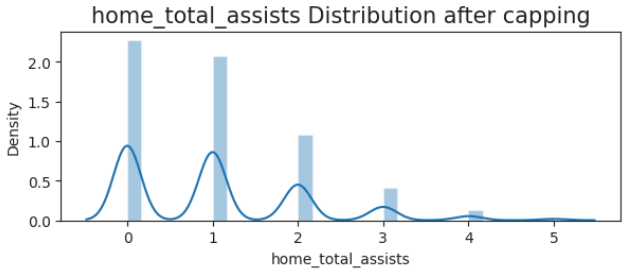


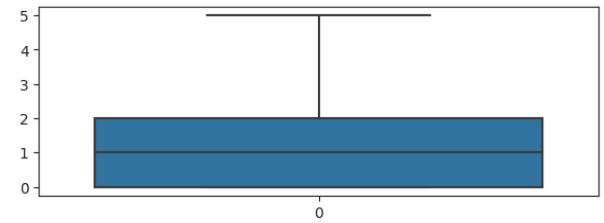

home_redCards Distribution after capping
Density
home_redCards

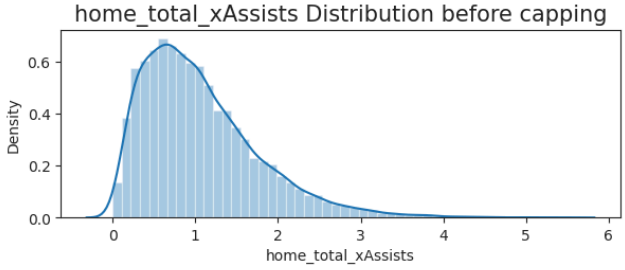


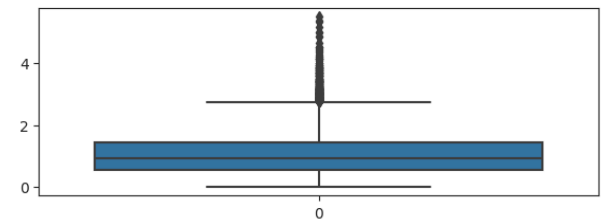

home_total_assists Distribution before capping
Density
home_total_assists

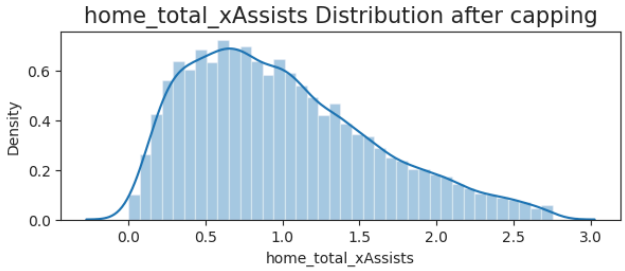


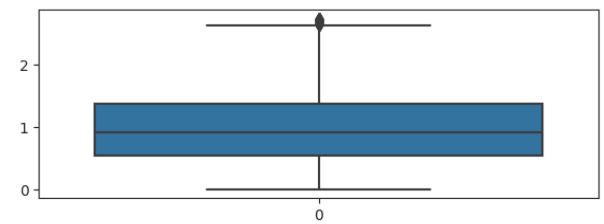

home_total_assists Distribution after capping
Density
home_total_assists

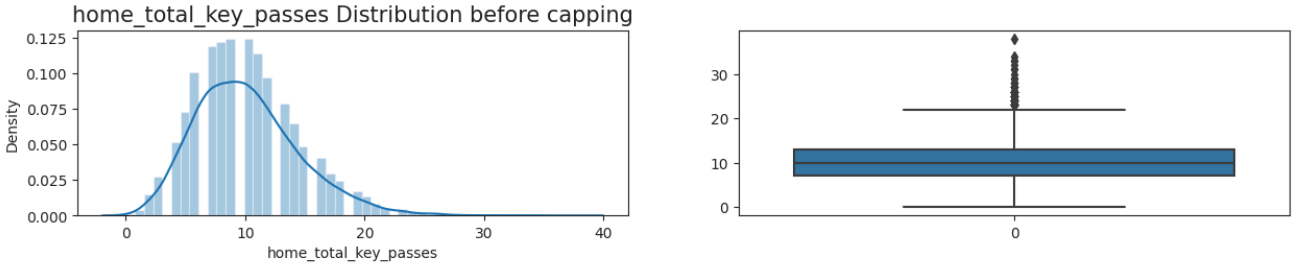


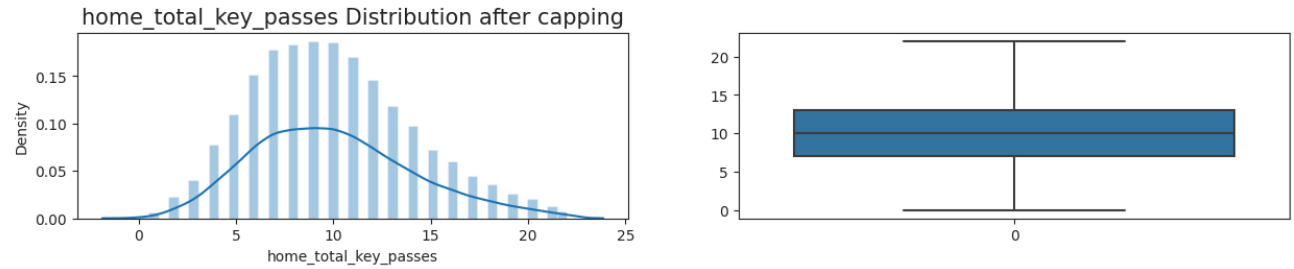

home_total_xAssists Distribution before capping
Density
home_total_xAssists

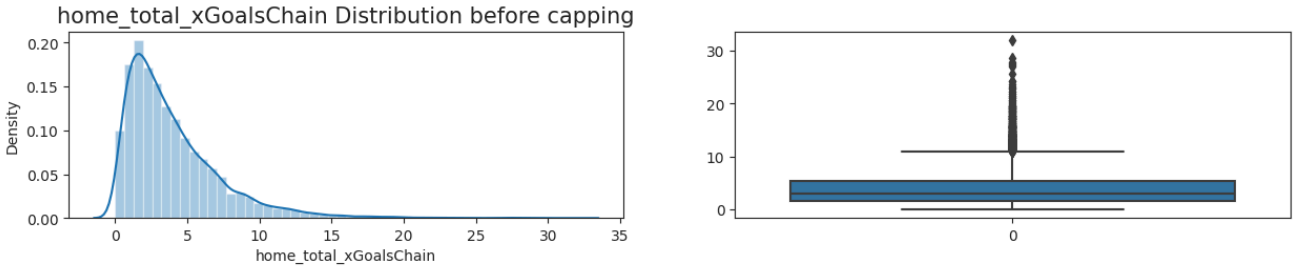


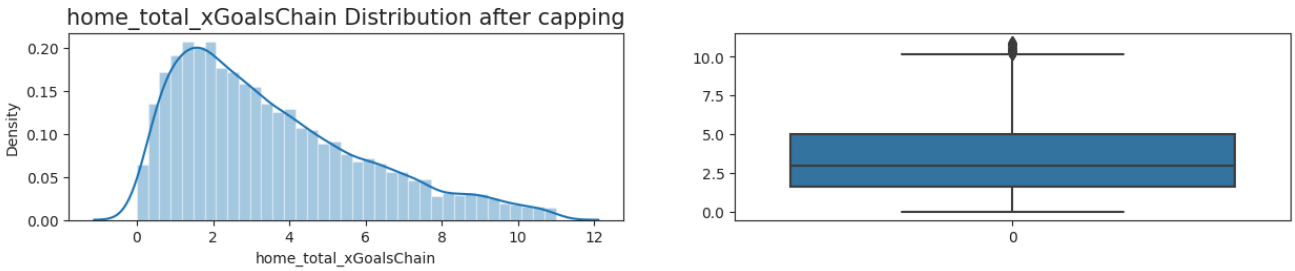

home_total_xAssists Distribution after capping
Density
home_total_xAssists

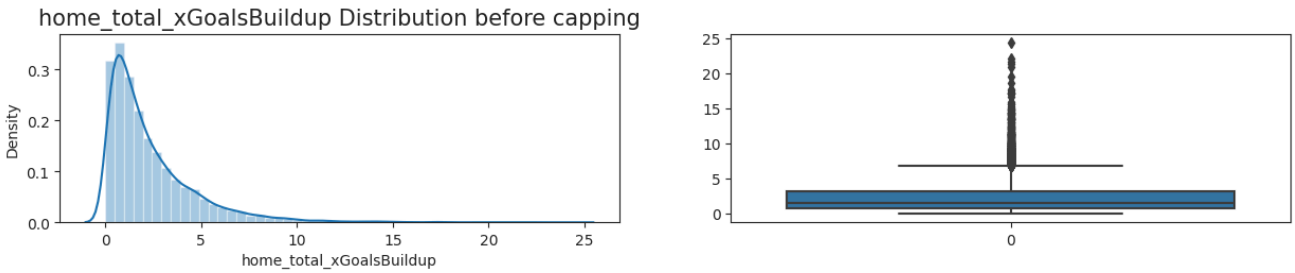


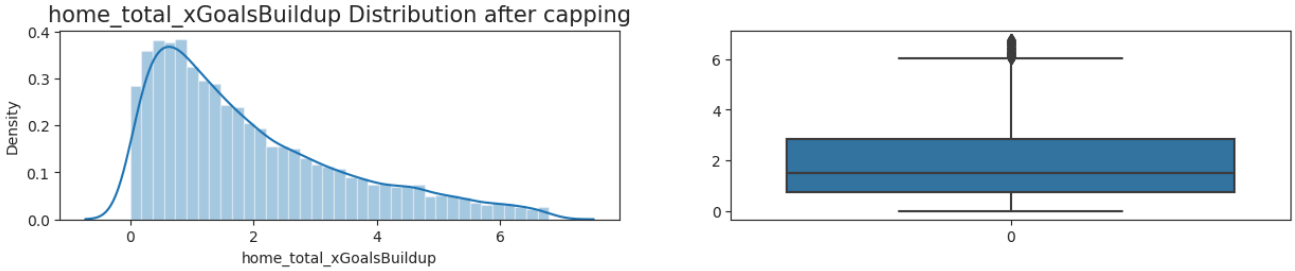

home_total_key_passes Distribution before capping
Density
home_total_key_passes

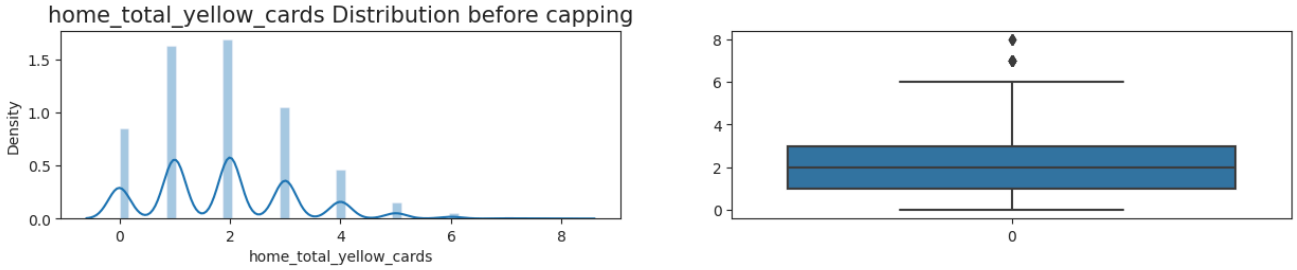

home_total_key_passes Distribution after capping
Density
home_total_key_passes

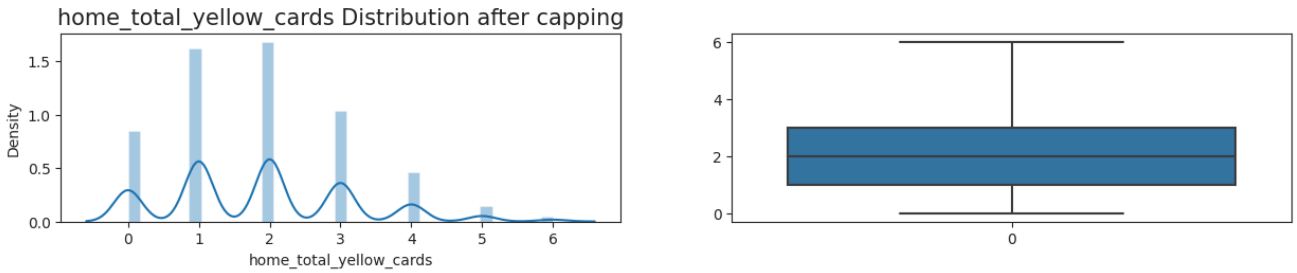

home_total_xGoalsChain Distribution before capping
Density
home_total_xGoalsChain

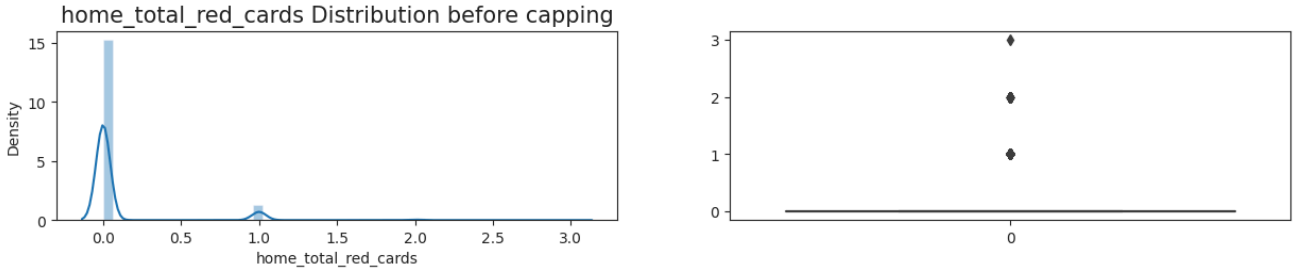

home_total_xGoalsChain Distribution after capping
Density
home_total_xGoalsChain

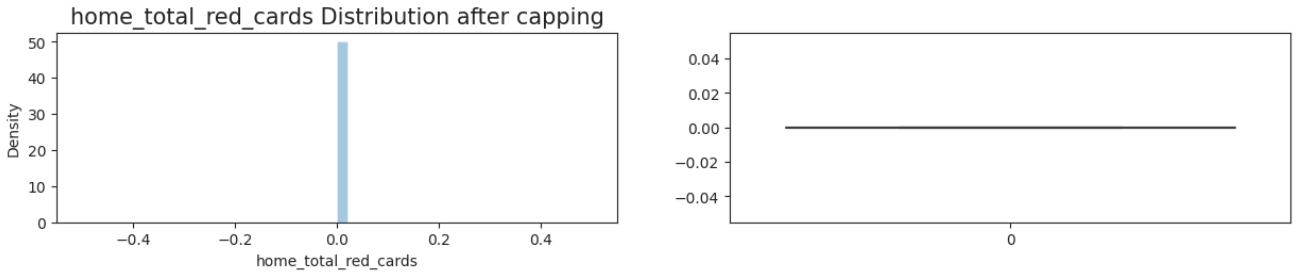

home_total_xGoalsBuildup Distribution before capping
Density
home_total_xGoalsBuildup

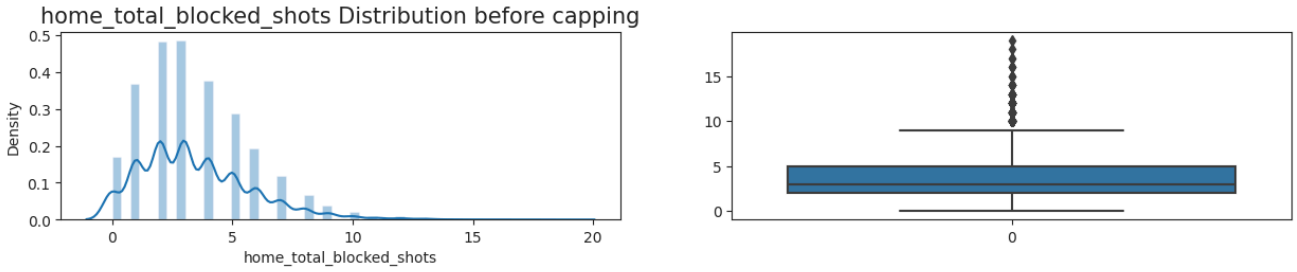

home_total_xGoalsBuildup Distribution after capping
Density
home_total_xGoalsBuildup

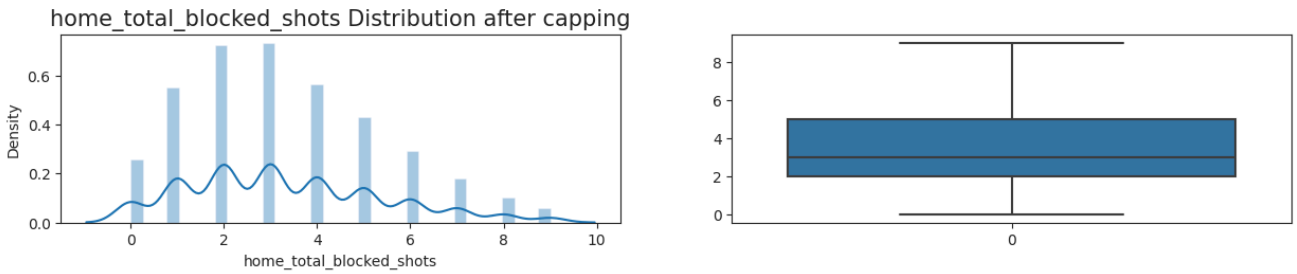

home_total_yellow_cards Distribution before capping
Density
home_total_yellow_cards

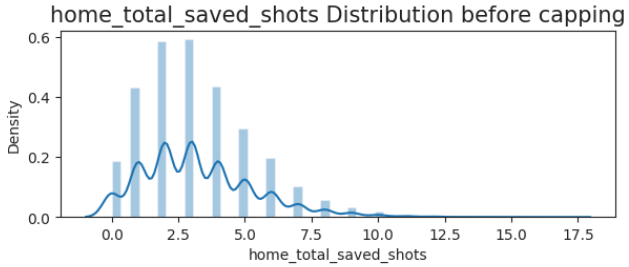

home_total_yellow_cards Distribution after capping
Density
home_total_yellow_cards

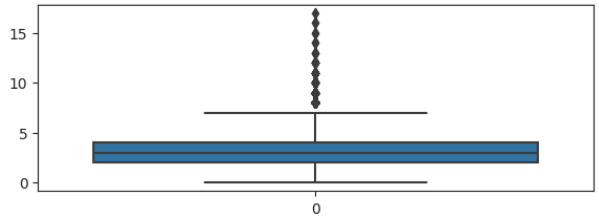

home_total_red_cards Distribution before capping
Density
home_total_red_cards

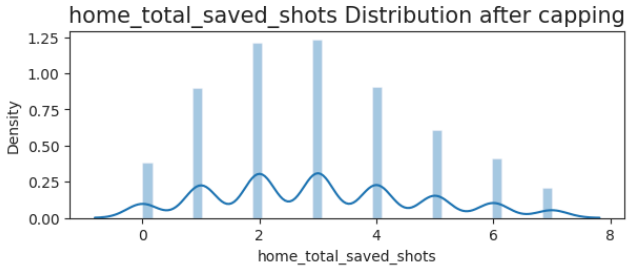

home_total_red_cards Distribution after capping
Density
home_total_red_cards

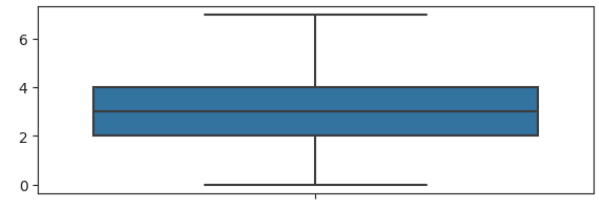

home_total_blocked_shots Distribution before capping
Density
home_total_blocked_shots

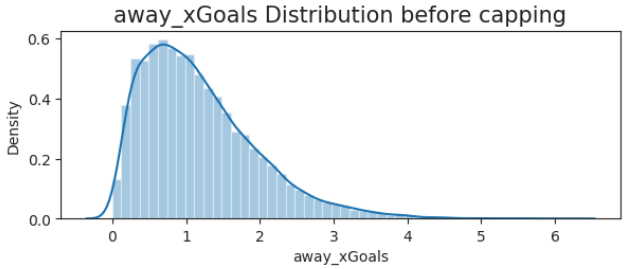

home_total_blocked_shots Distribution after capping
Density
home_total_blocked_shots

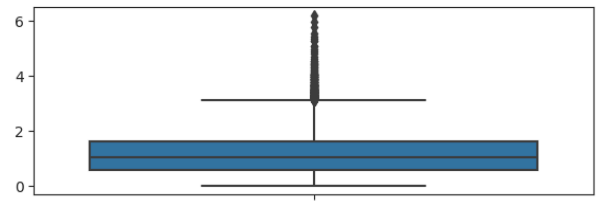

home_total_saved_shots Distribution before capping
Density
home_total_saved_shots

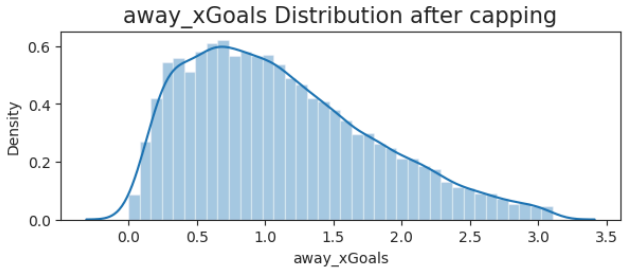


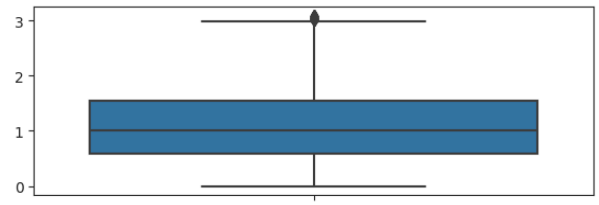

home_total_saved_shots Distribution after capping
Density
home_total_saved_shots

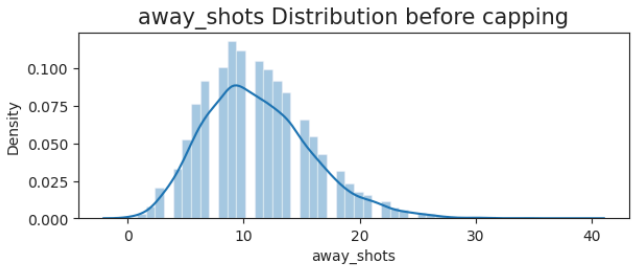


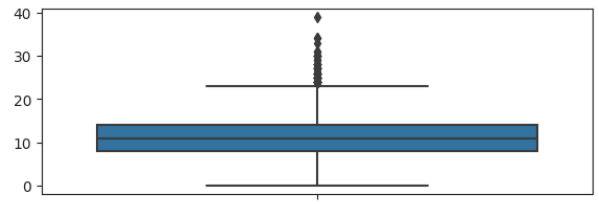

away_xGoals Distribution before capping
Density
away_xGoals

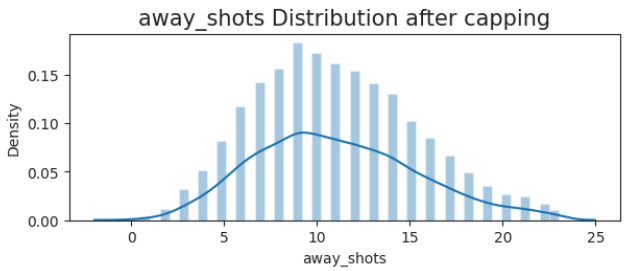


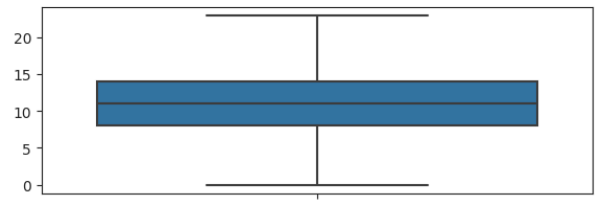

away_xGoals Distribution after capping
Density
away_xGoals

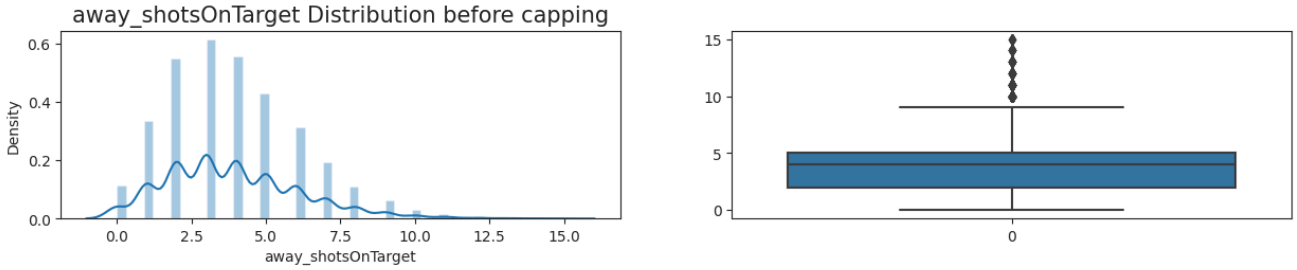


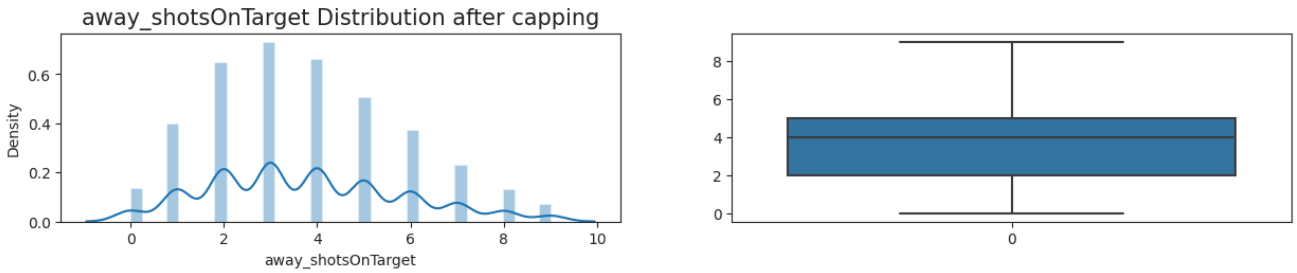

away_shots Distribution before capping
Density
away_shots

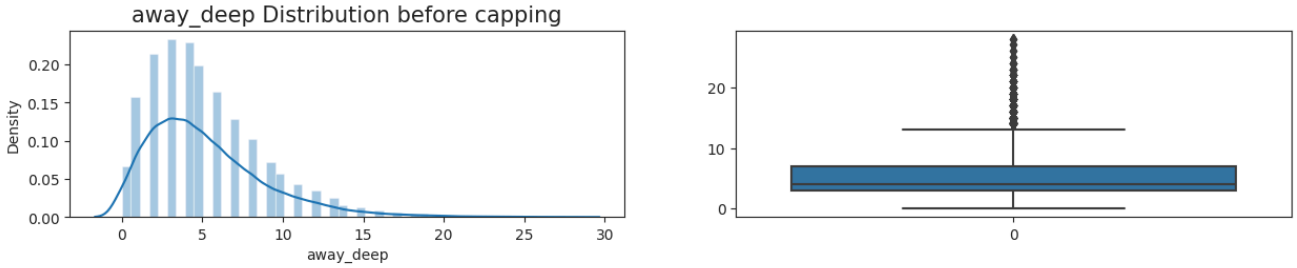


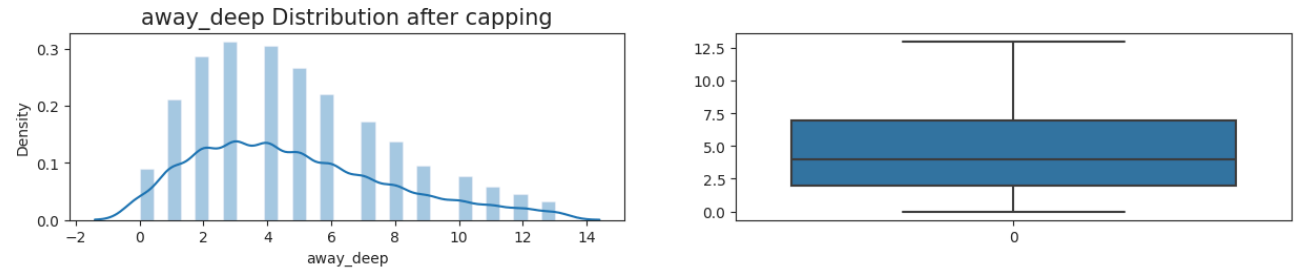

away_shots Distribution after capping
Density
away_shots

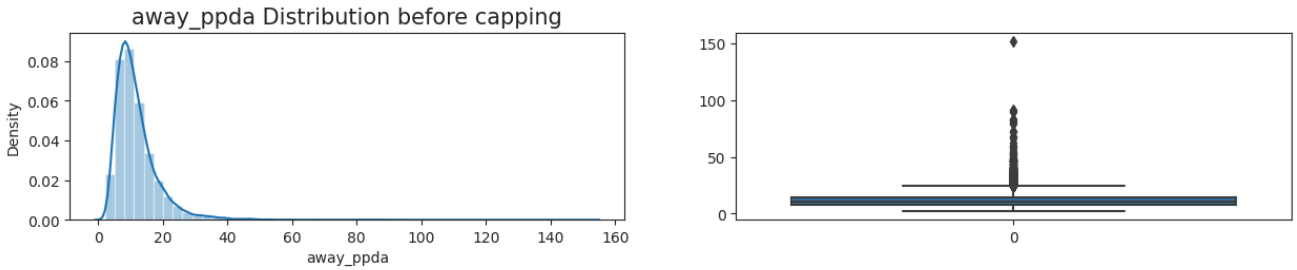


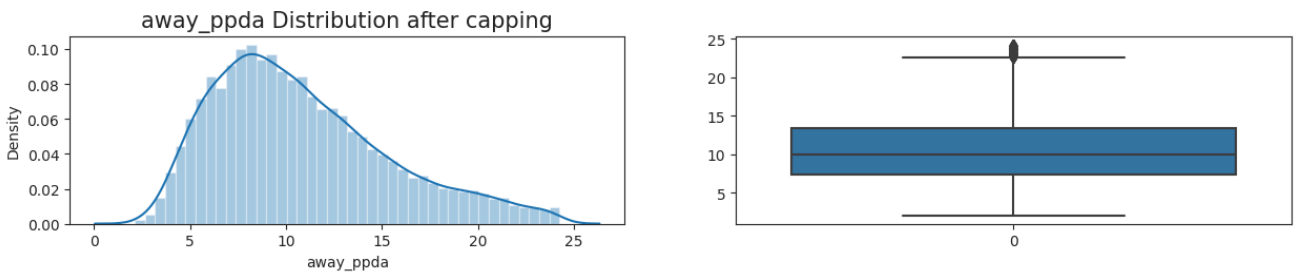

away_shotsOnTarget Distribution before capping
Density
away_shotsOnTarget

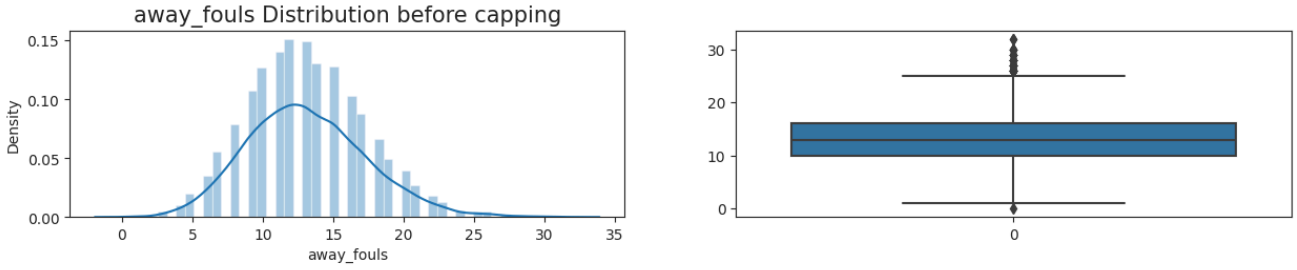

away_shotsOnTarget Distribution after capping
Density
away_shotsOnTarget

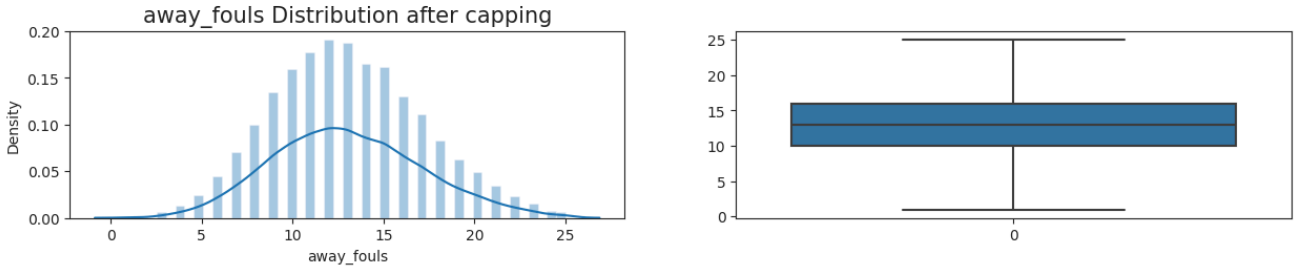

away_deep Distribution before capping
Density
away_deep

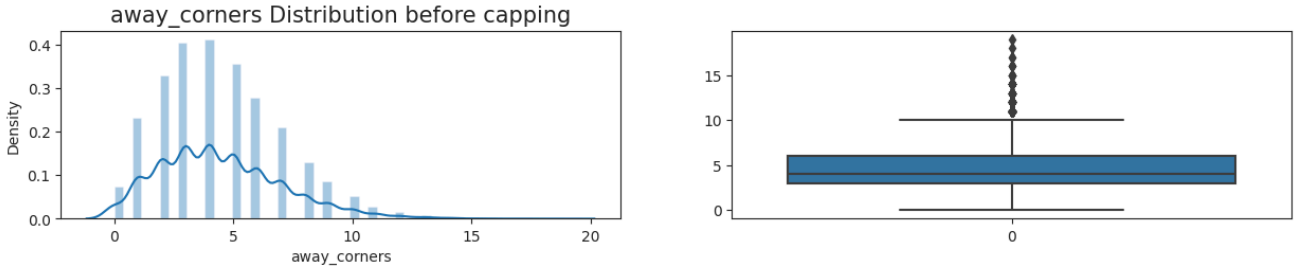

away_deep Distribution after capping
Density
away_deep

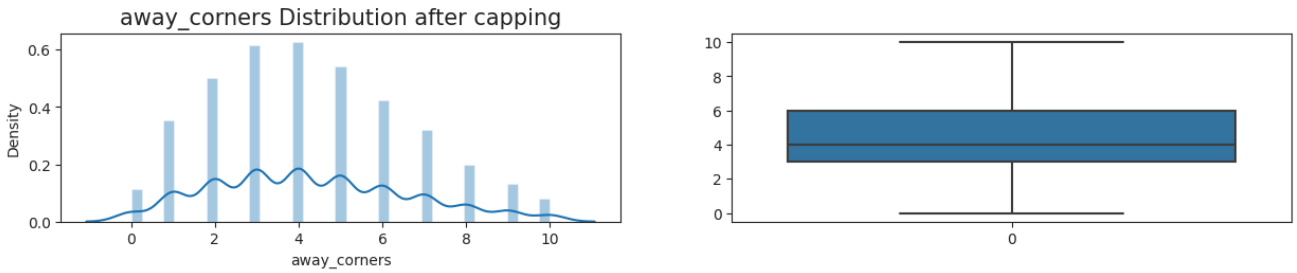

away_ppda Distribution before capping
Density
away_ppda

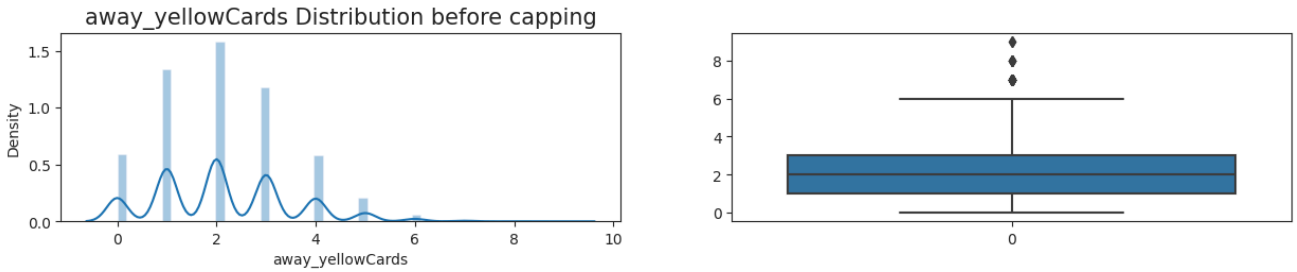

away_ppda Distribution after capping
Density
away_ppda

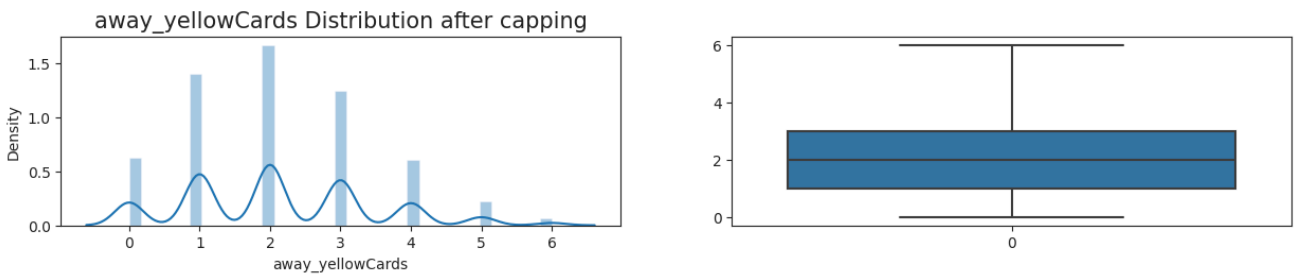

away_fouls Distribution before capping
Density
away_fouls

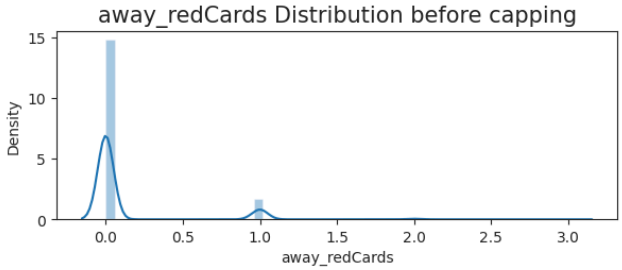

away_fouls Distribution after capping
Density
away_fouls

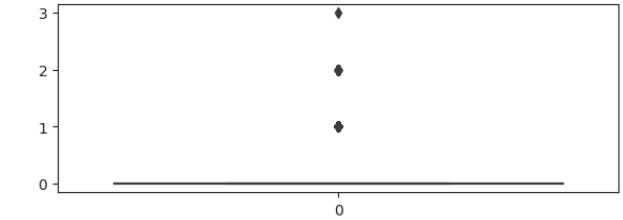

away_corners Distribution before capping
Density
away_corners

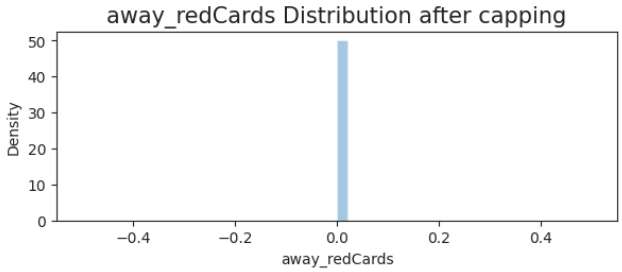

away_corners Distribution after capping
Density
away_corners

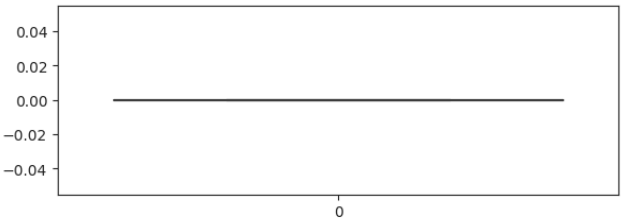

away_yellowCards Distribution before capping
Density
away_yellowCards

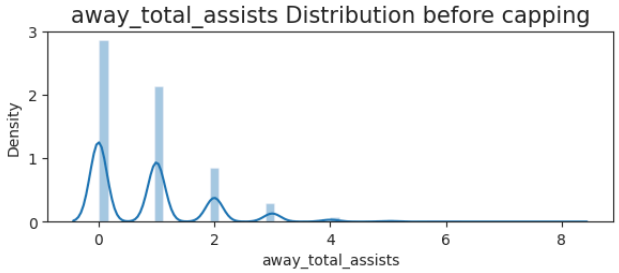

away_yellowCards Distribution after capping
Density
away_yellowCards

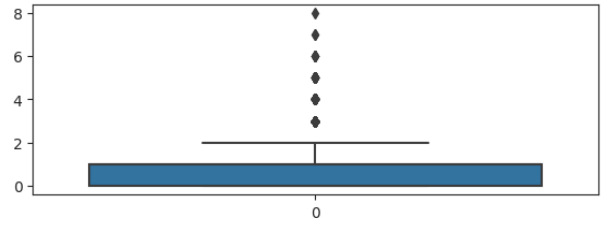

away_redCards Distribution before capping
Density
away_redCards

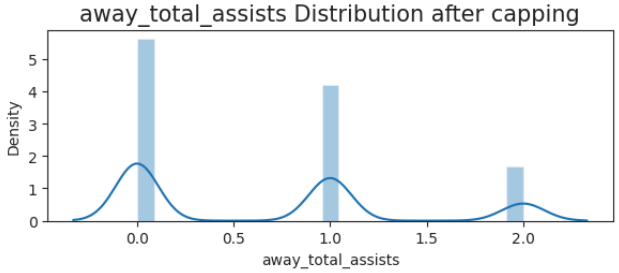


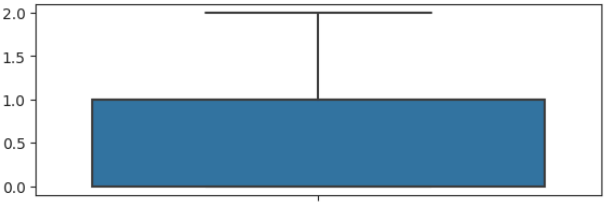

away_redCards Distribution after capping
Density
away_redCards

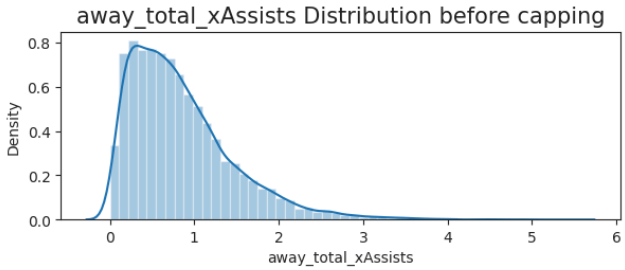


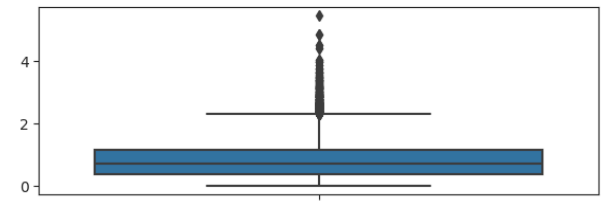

away_total_assists Distribution before capping
Density
away_total_assists

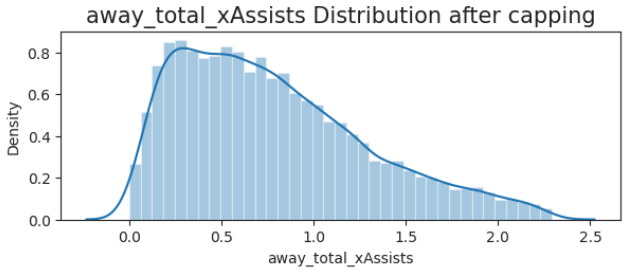


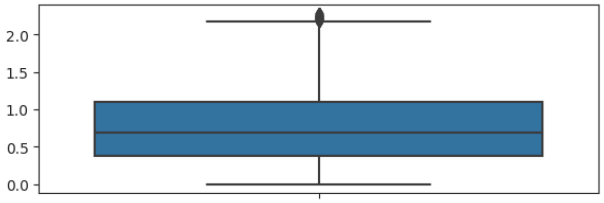

away_total_assists Distribution after capping
Density
away_total_assists

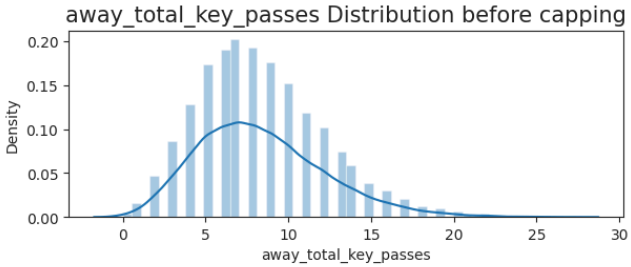


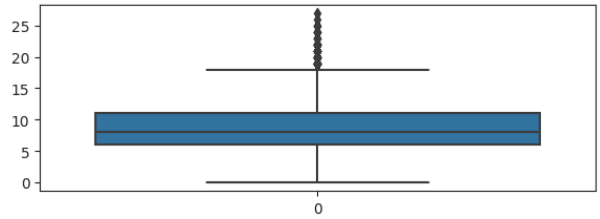

away_total_xAssists Distribution before capping
Density
away_total_xAssists

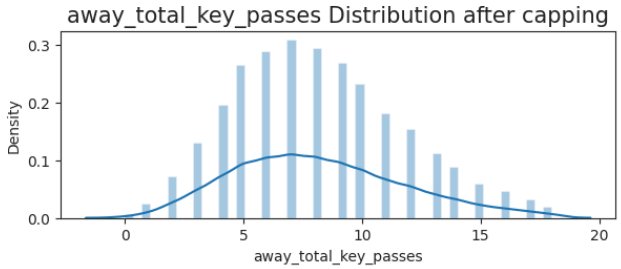


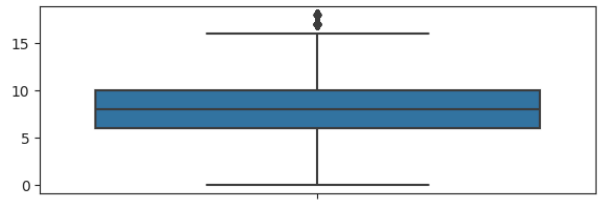

away_total_xAssists Distribution after capping
Density
away_total_xAssists

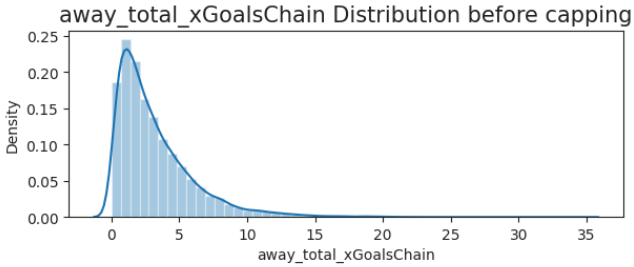


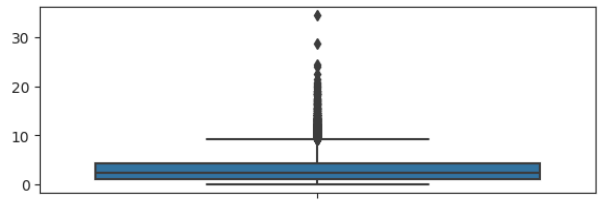

away_total_key_passes Distribution before capping
Density
away_total_key_passes

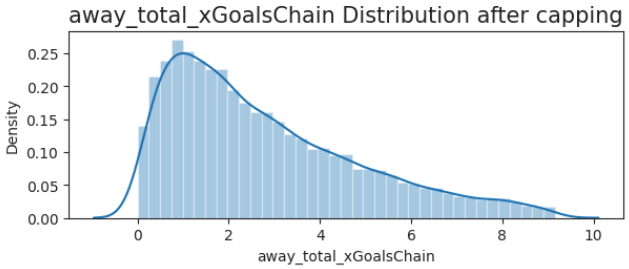


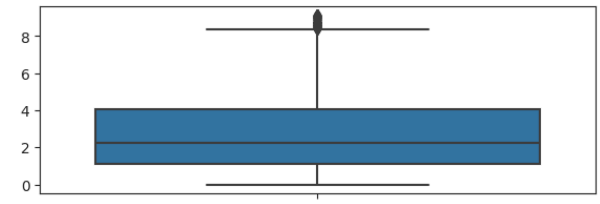

away_total_key_passes Distribution after capping
Density
away_total_key_passes

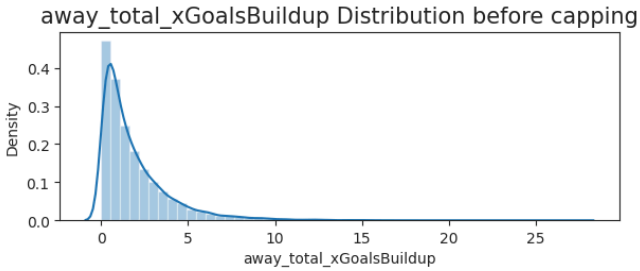


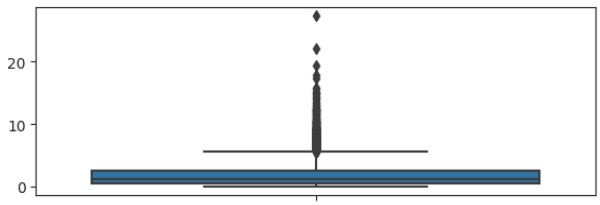

away_total_xGoalsChain Distribution before capping
Density
away_total_xGoalsChain

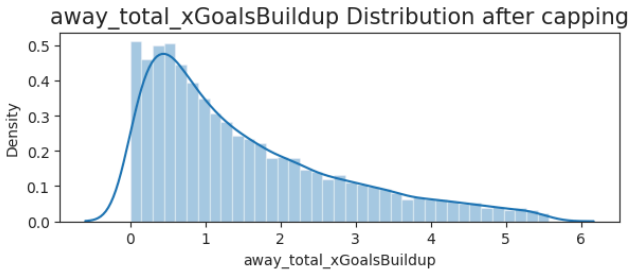


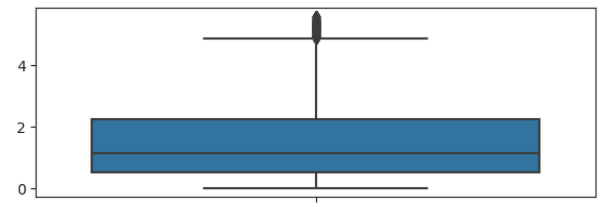

away_total_xGoalsChain Distribution after capping
Density
away_total_xGoalsChain

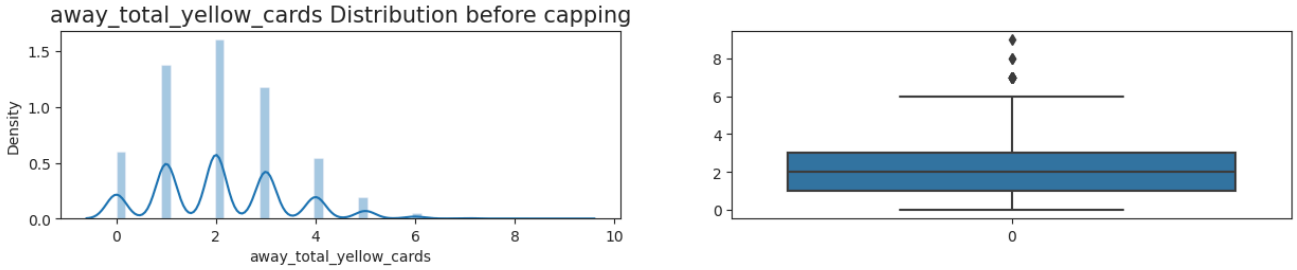


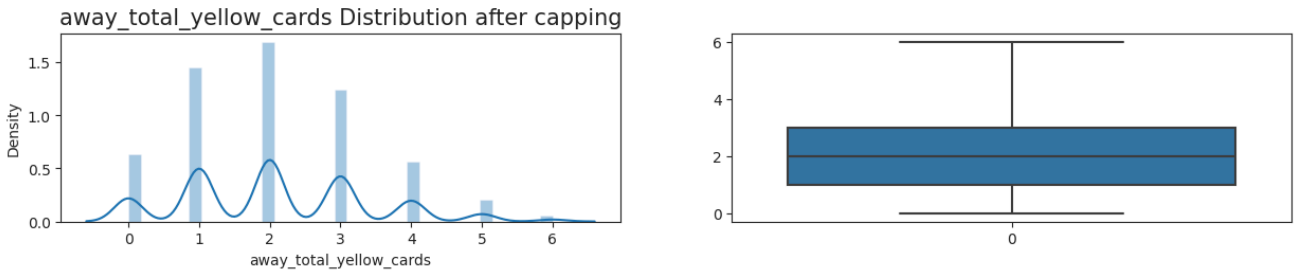

away_total_xGoalsBuildup Distribution before capping
Density
away_total_xGoalsBuildup

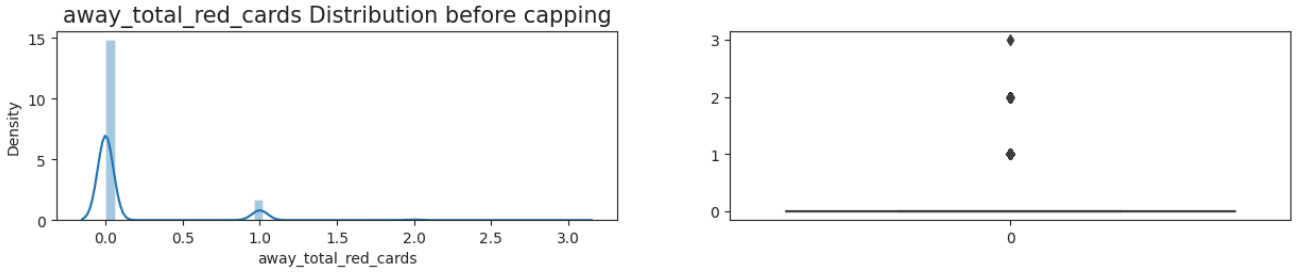


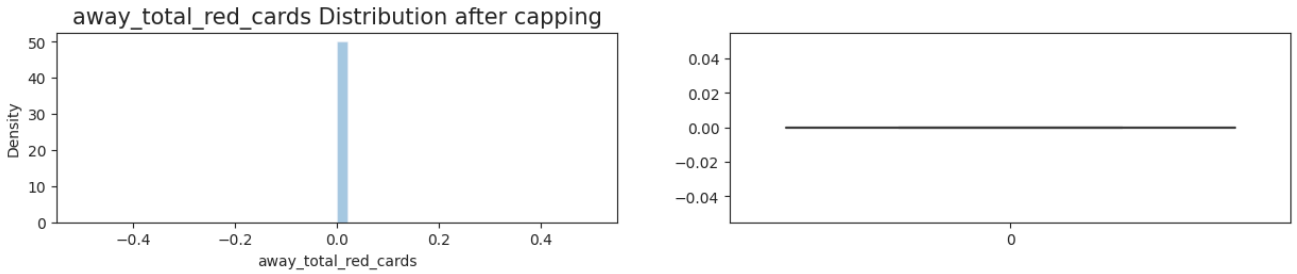

away_total_xGoalsBuildup Distribution after capping
Density
away_total_xGoalsBuildup

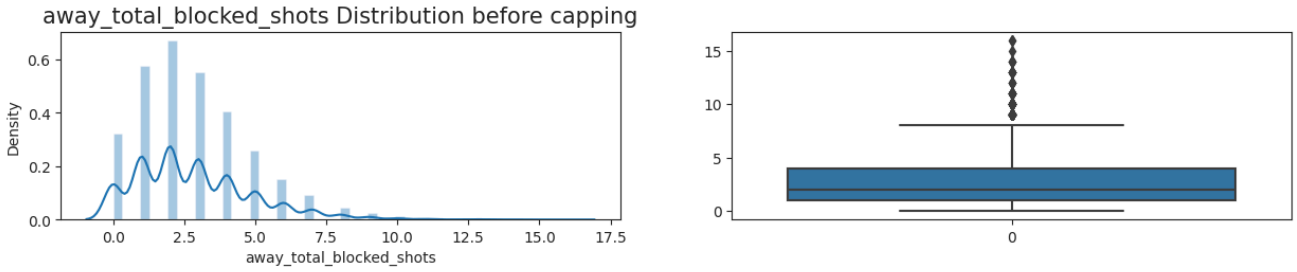


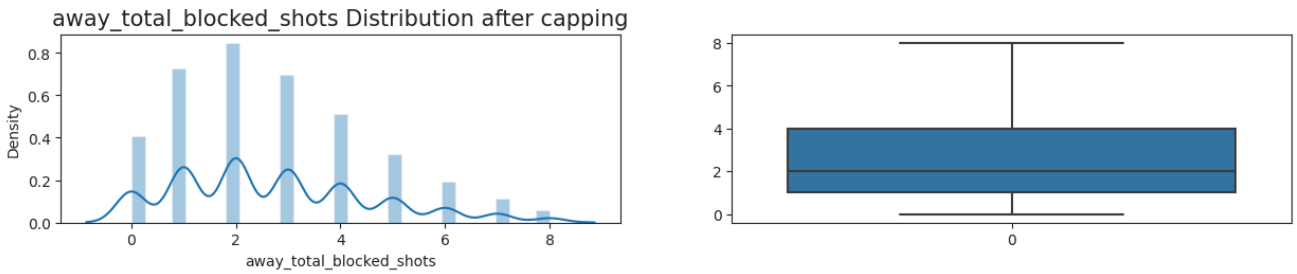

away_total_yellow_cards Distribution before capping
Density
away_total_yellow_cards

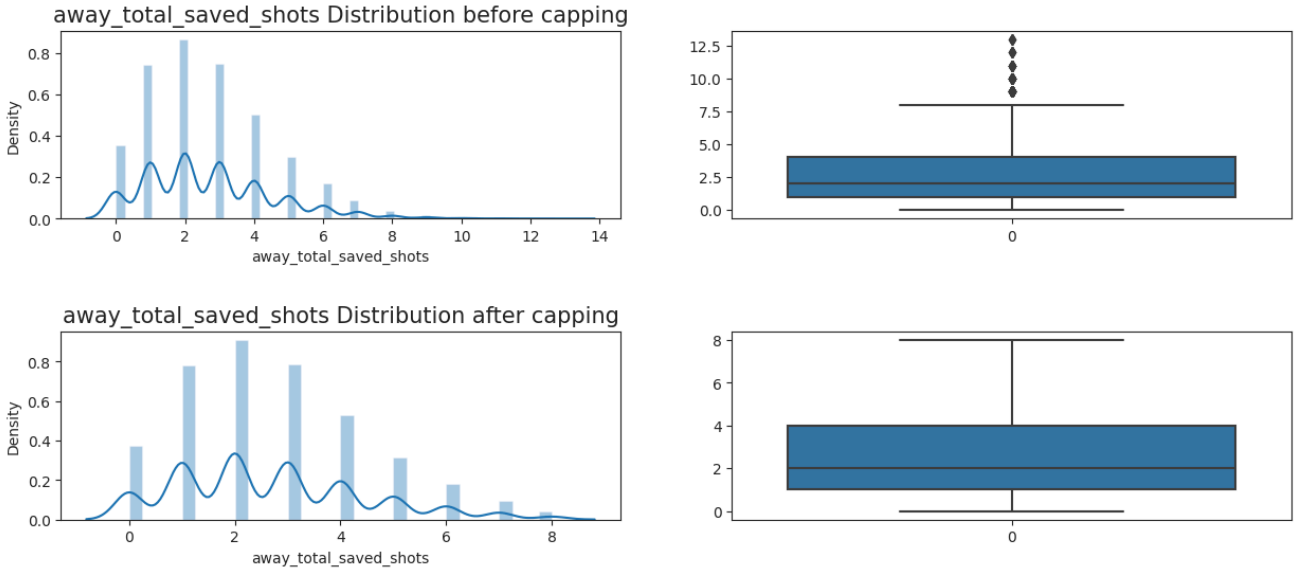

away_total_yellow_cards Distribution after capping
Density
away_total_yellow_cards


away_total_red_cards Distribution before capping
Density
away_total_red_cards


away_total_red_cards Distribution after capping
Density
away_total_red_cards


away_total_blocked_shots Distribution before capping
Density
away_total_blocked_shots


away_total_blocked_shots Distribution after capping
Density
away_total_blocked_shots



We clearly need to filter some features that either the feature distribution has changed significantly, or cases where removing outliers results with only one value.

Checking if distribution (Kolmogorov–Smirnov) and/or correlation has changed significantly:

	feature	outliers_cnt	distribution_changed	correlation_changed
0	home_Goals	981	+	+
1	away_Goals	44	-	-
2	home_GoalsHalfTime	403	+	+
3	away_GoalsHalfTime	225	+	+
4	home_xGoals	276	+	-
5	home_shots	167	-	-
6	home_shotsOnTarget	351	+	-
7	home_deep	227	+	-
8	home_ppda	540	+	+
9	home_fouls	241	+	-
10	home_corners	143	-	-
11	home_yellowCards	37	-	-
12	home_redCards	1078	+	-
13	home_total_assists	20	-	-
14	home_total_xAssists	346	+	-
15	home_total_key_passes	123	-	-
16	home_total_xGoalsChain	523	+	+
17	home_total_xGoalsBuildup	664	+	+
18	home_total_yellow_cards	20	-	-
19	home_total_red_cards	1064	+	-
20	home_total_blocked_shots	230	+	-
21	home_total_saved_shots	532	+	-
22	away_xGoals	312	+	+
23	away_shots	161	-	-
24	away_shotsOnTarget	233	+	-
25	away_deep	423	+	-
26	away_ppda	606	+	+
27	away_fouls	81	-	-
28	away_corners	283	+	-
29	away_yellowCards	43	-	-
30	away_redCards	1396	+	-
31	away_total_assists	790	+	+
32	away_total_xAssists	389	+	+
33	away_total_key_passes	169	-	-
34	away_total_xGoalsChain	538	+	+
35	away_total_xGoalsBuildup	718	+	+

	feature	outliers_cnt	distribution_changed	correlation_changed
36	away_total_yellow_cards	26	-	-
37	away_total_red_cards	1382	+	-
38	away_total_blocked_shots	187	-	-
39	away_total_saved_shots	95	-	-

	feature	outliers_cnt	distribution_changed	correlation_changed	drop
0	home_Goals	981	+	+	no
1	away_Goals	44	-	-	yes
2	home_GoalsHalfTime	403	+	+	no
3	away_GoalsHalfTime	225	+	+	no
4	home_xGoals	276	+	-	yes
5	home_shots	167	-	-	yes
6	home_shotsOnTarget	351	+	-	yes
7	home_deep	227	+	-	yes
8	home_ppda	540	+	+	no
9	home_fouls	241	+	-	yes
10	home_corners	143	-	-	yes
11	home_yellowCards	37	-	-	yes
12	home_redCards	1078	+	-	yes
13	home_total_assists	20	-	-	yes
14	home_total_xAssists	346	+	-	yes
15	home_total_key_passes	123	-	-	yes
16	home_total_xGoalsChain	523	+	+	no
17	home_total_xGoalsBuildup	664	+	+	no
18	home_total_yellow_cards	20	-	-	yes
19	home_total_red_cards	1064	+	-	yes
20	home_total_blocked_shots	230	+	-	yes
21	home_total_saved_shots	532	+	-	yes
22	away_xGoals	312	+	+	no
23	away_shots	161	-	-	yes
24	away_shotsOnTarget	233	+	-	yes
25	away_deep	423	+	-	yes
26	away_ppda	606	+	+	no
27	away_fouls	81	-	-	yes
28	away_corners	283	+	-	yes
29	away_yellowCards	43	-	-	yes
30	away_redCards	1396	+	-	yes
31	away_total_assists	790	+	+	no
32	away_total_xAssists	389	+	+	no
33	away_total_key_passes	169	-	-	yes
34	away_total_xGoalsChain	538	+	+	no
35	away_total_xGoalsBuildup	718	+	+	no

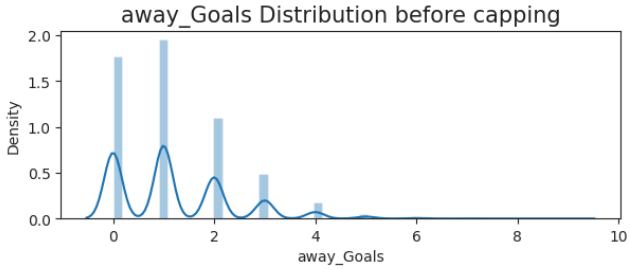
	feature	outliers_cnt	distribution_changed	correlation_changed	drop
36	away_total_yellow_cards	26	-	-	yes
37	away_total_red_cards	1382	+	-	yes
38	away_total_blocked_shots	187	-	-	yes
39	away_total_saved_shots	95	-	-	yes

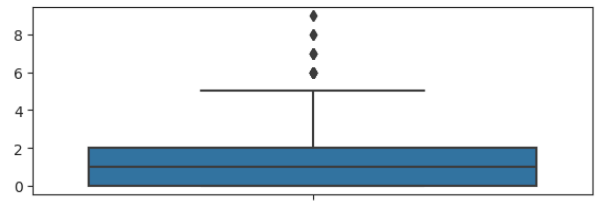
```
drop
no     12
yes    28
dtype: int64
```

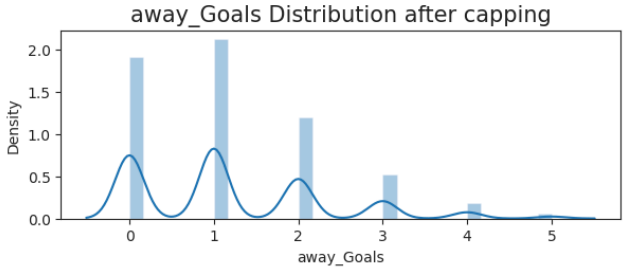
There are 28 features in total where removal of the outliers does not affect neither correlation nor distribution. And 40 features that shall be removed.

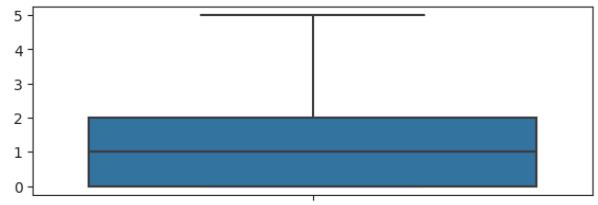
```
1                   away_Goals
4                  home_xGoals
5                   home_shots
6           home_shotsOnTarget
7                    home_deep
9                   home_fouls
10                home_corners
11             home_yellowCards
12                home_redCards
13           home_total_assists
14          home_total_xAssists
15        home_total_key_passes
18      home_total_yellow_cards
19         home_total_red_cards
20     home_total_blocked_shots
21       home_total_saved_shots
23                  away_shots
24          away_shotsOnTarget
25                   away_deep
27                  away_fouls
28                away_corners
29             away_yellowCards
30                away_redCards
33        away_total_key_passes
36      away_total_yellow_cards
37         away_total_red_cards
38     away_total_blocked_shots
39       away_total_saved_shots
Name: feature, dtype: object
```

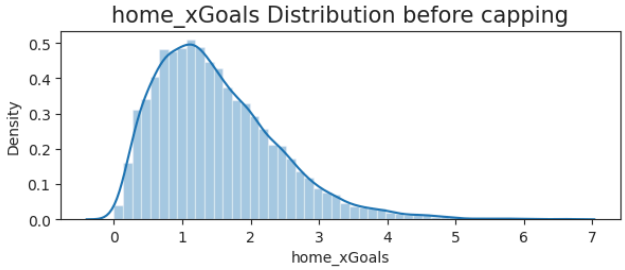
Let's see the distribution with and without outliers for does specific 15 features:

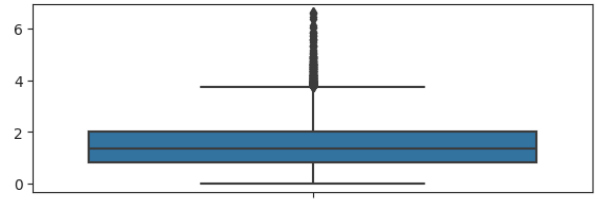

away_Goals Distribution before capping
Density
away_Goals

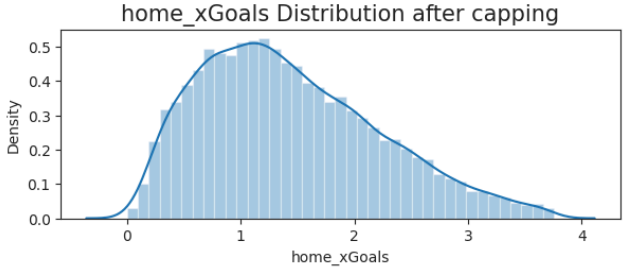


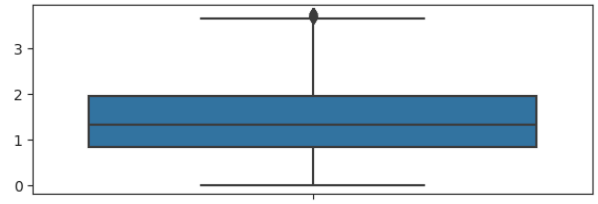

away_Goals Distribution after capping
Density
away_Goals

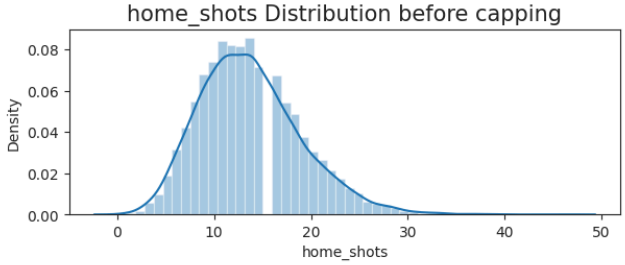


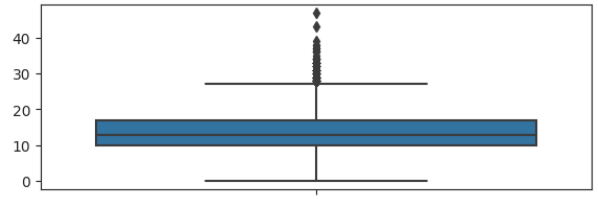

home_xGoals Distribution before capping
Density
home_xGoals

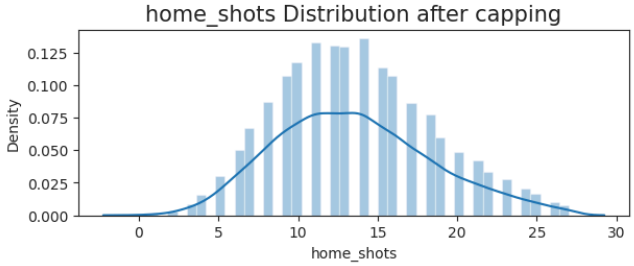


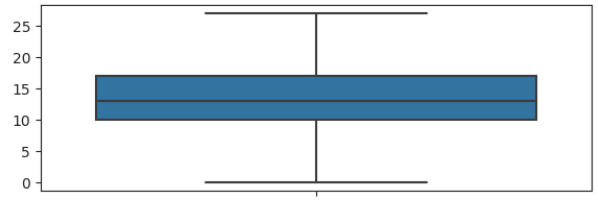

home_xGoals Distribution after capping
Density
home_xGoals

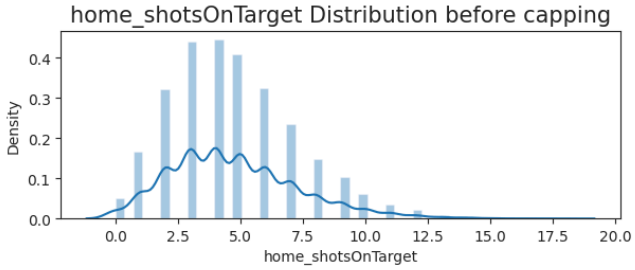


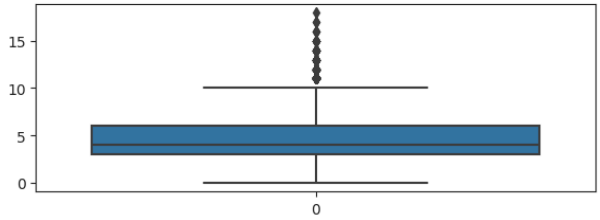

home_shots Distribution before capping
Density
home_shots

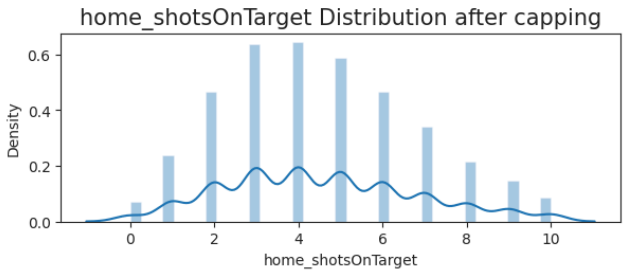


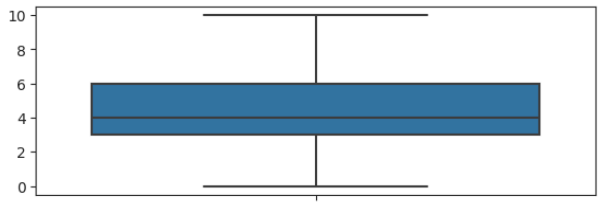

home_shots Distribution after capping
Density
home_shots

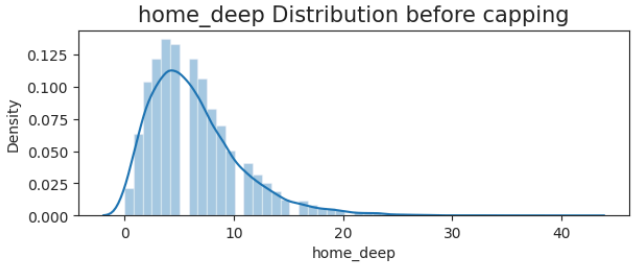


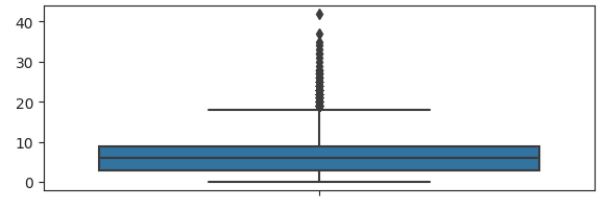

home_shotsOnTarget Distribution before capping
Density
home_shotsOnTarget

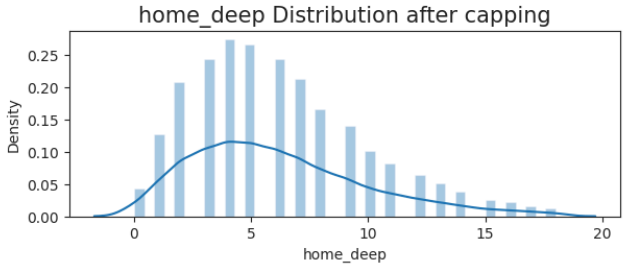


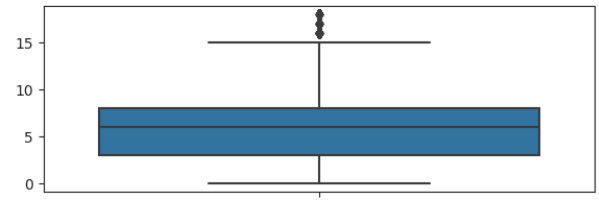

home_shotsOnTarget Distribution after capping
Density
home_shotsOnTarget

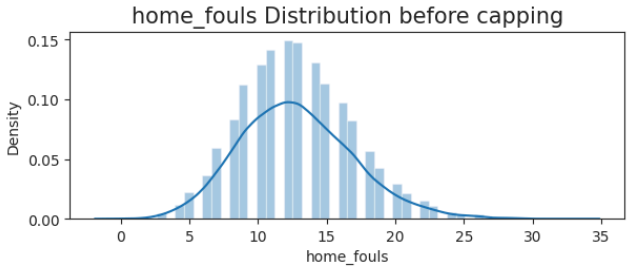


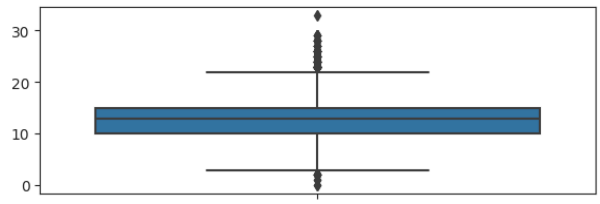

home_deep Distribution before capping
Density
home_deep

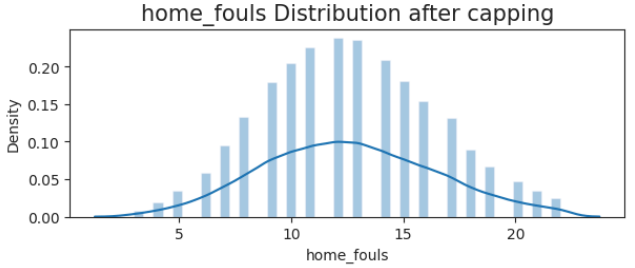


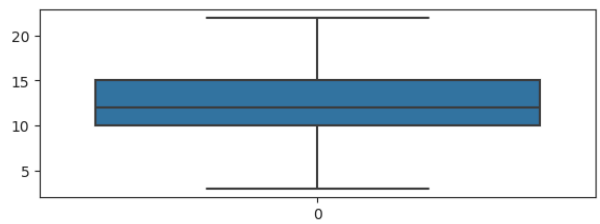

home_deep Distribution after capping
Density
home_deep

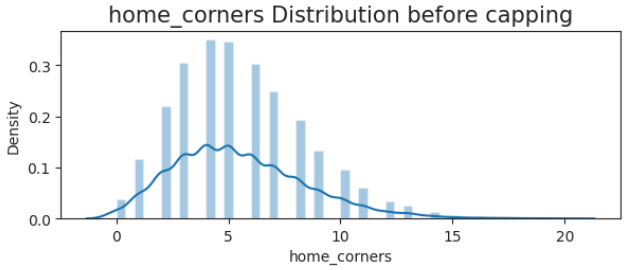


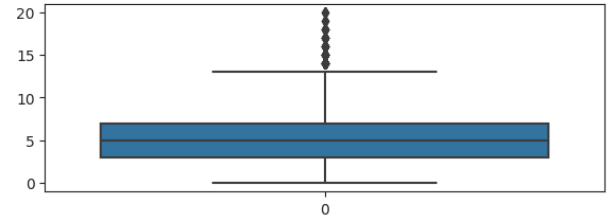

home_fouls Distribution before capping
Density
home_fouls

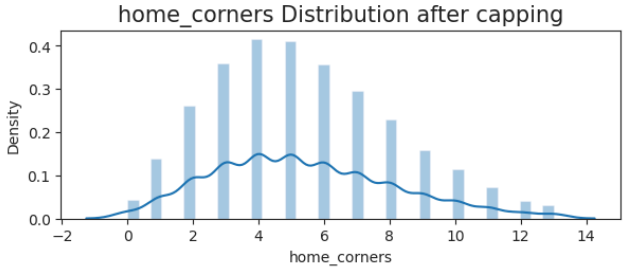


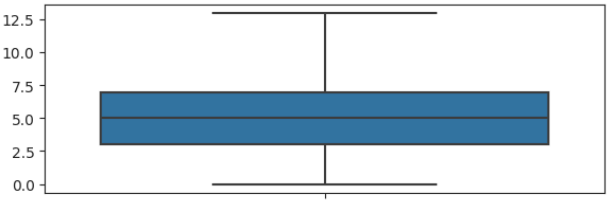

home_fouls Distribution after capping
Density
home_fouls

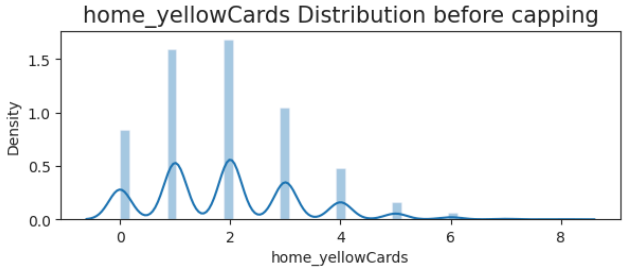


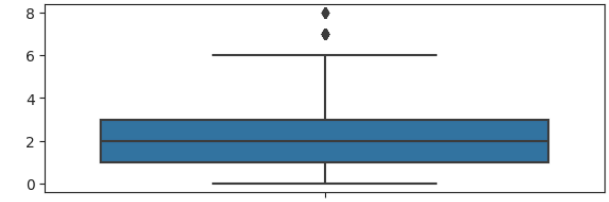

home_corners Distribution before capping
Density
home_corners

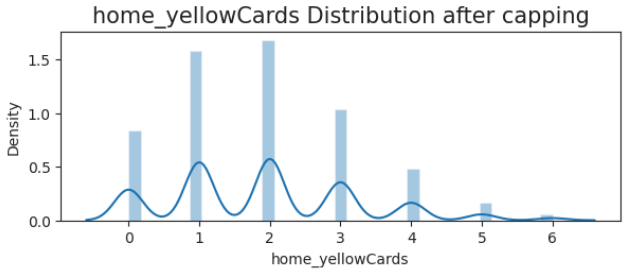


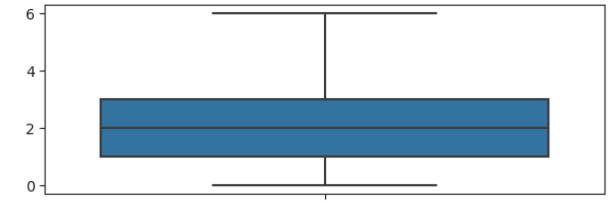

home_corners Distribution after capping
Density
home_corners

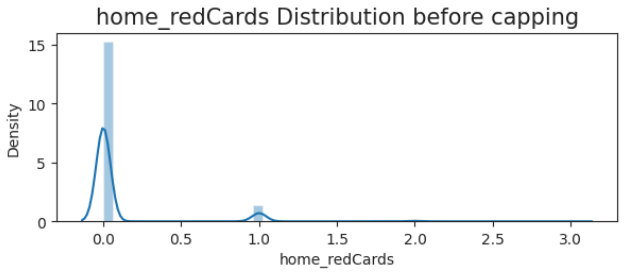


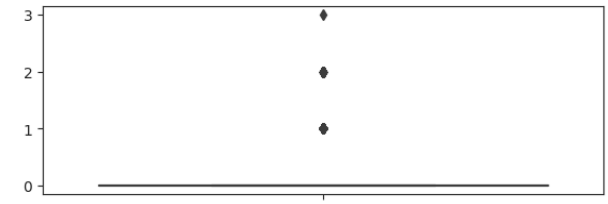

home_yellowCards Distribution before capping
Density
home_yellowCards

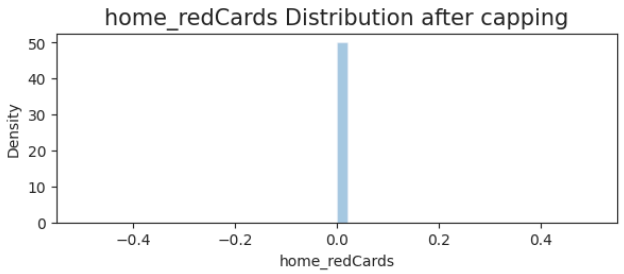


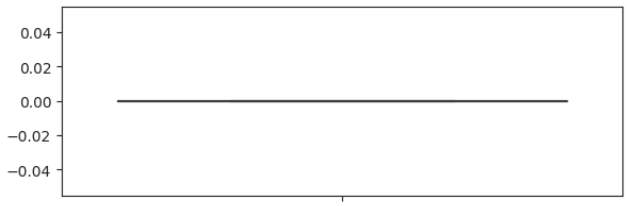

home_yellowCards Distribution after capping
Density
home_yellowCards

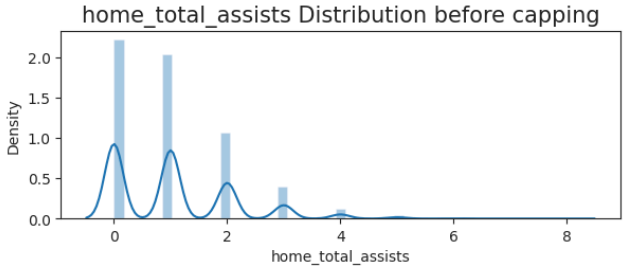


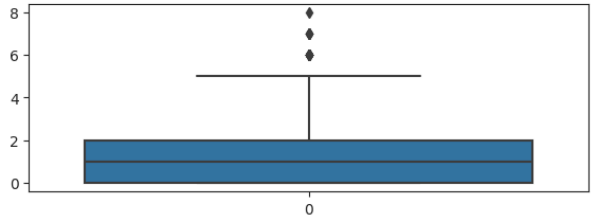

home_redCards Distribution before capping
Density
home_redCards

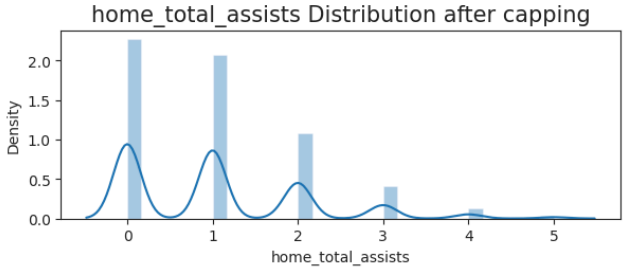


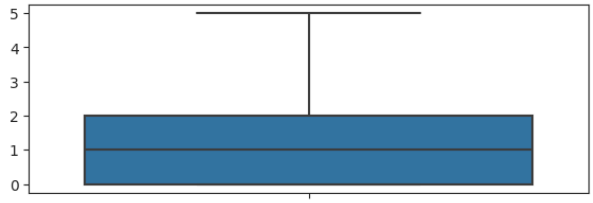

home_redCards Distribution after capping
Density
home_redCards

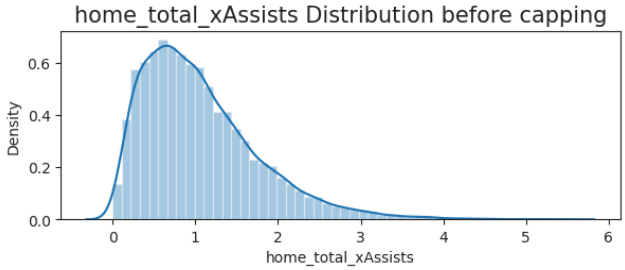


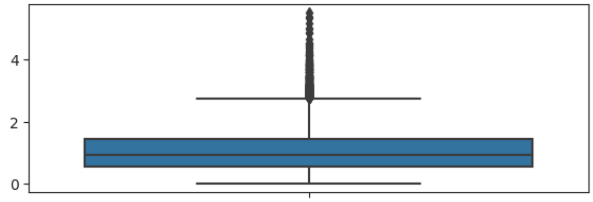

home_total_assists Distribution before capping
Density
home_total_assists

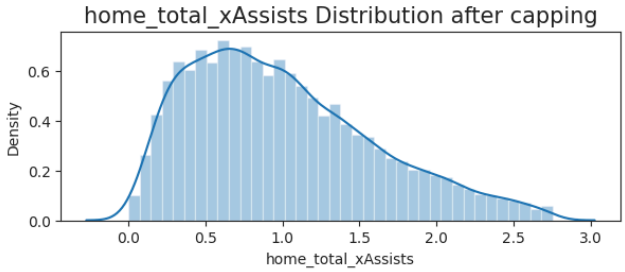


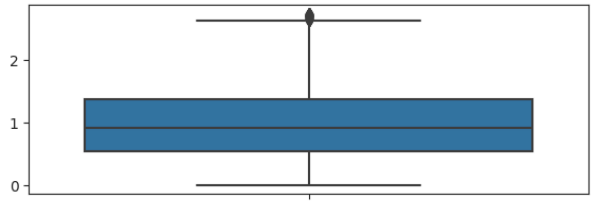

home_total_assists Distribution after capping
Density
home_total_assists

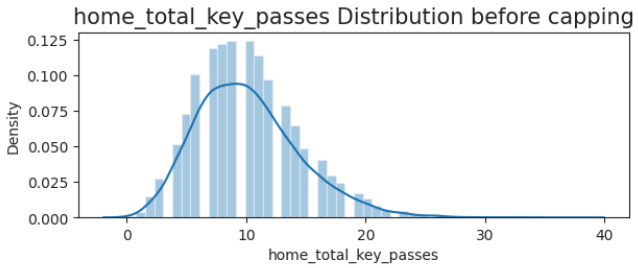


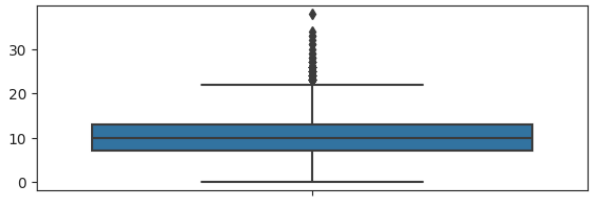

home_total_xAssists Distribution before capping
Density
home_total_xAssists

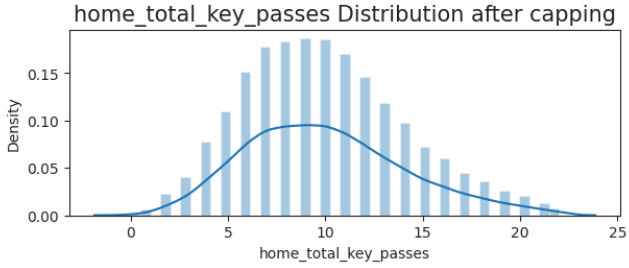


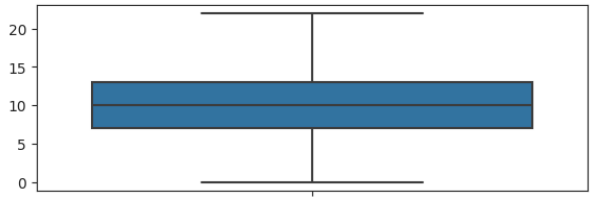

home_total_xAssists Distribution after capping
Density
home_total_xAssists

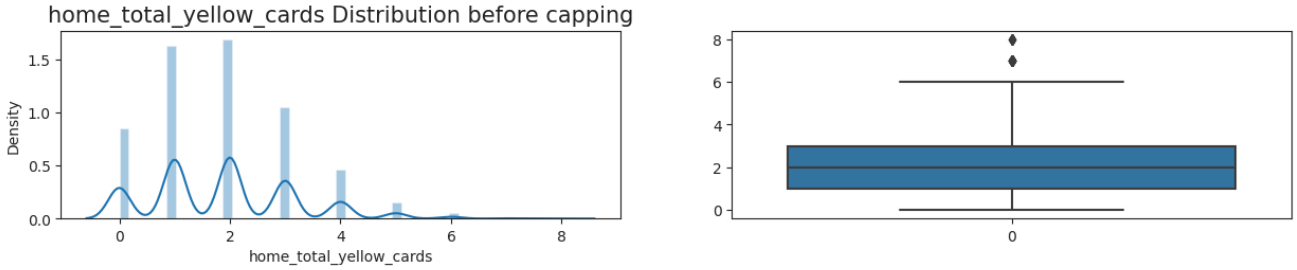


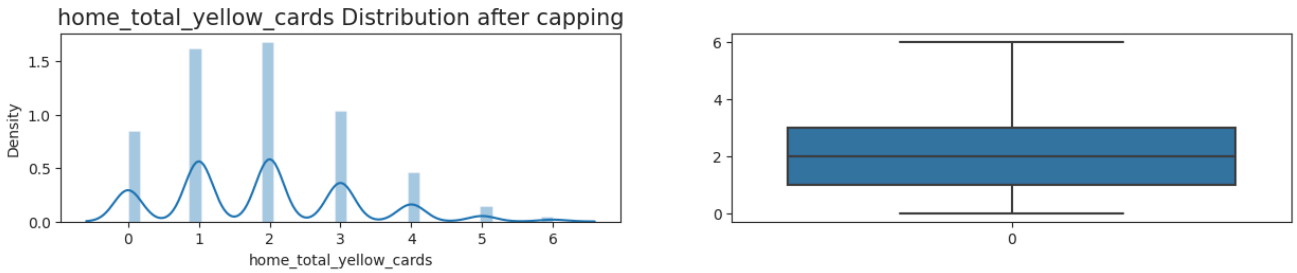

home_total_key_passes Distribution before capping
Density
home_total_key_passes

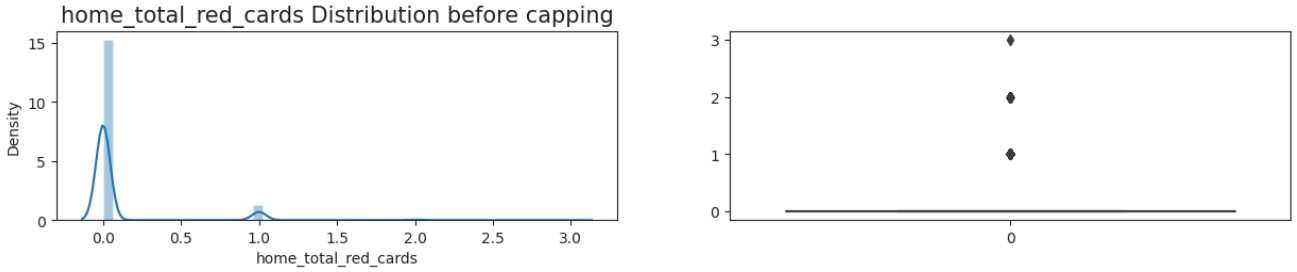


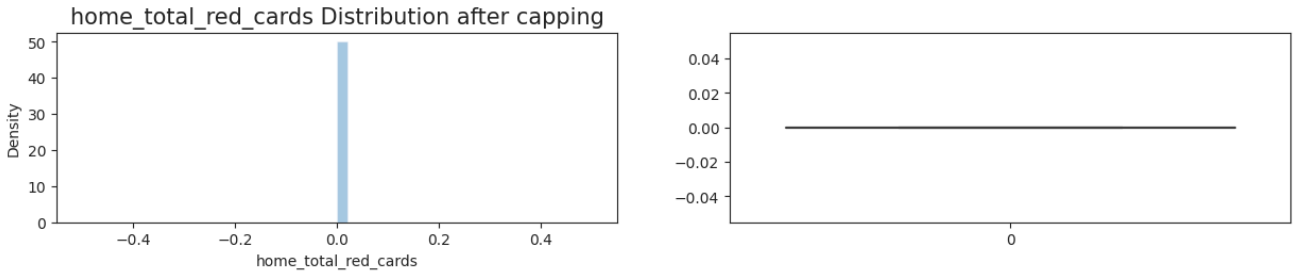

home_total_key_passes Distribution after capping
Density
home_total_key_passes

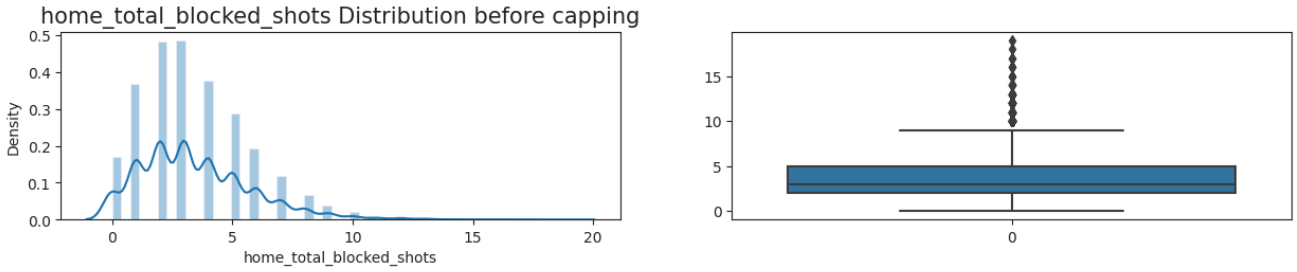


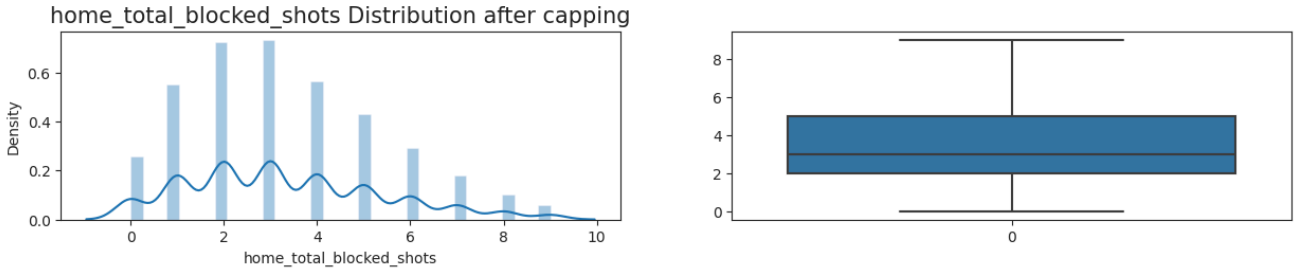

home_total_yellow_cards Distribution before capping
Density
home_total_yellow_cards

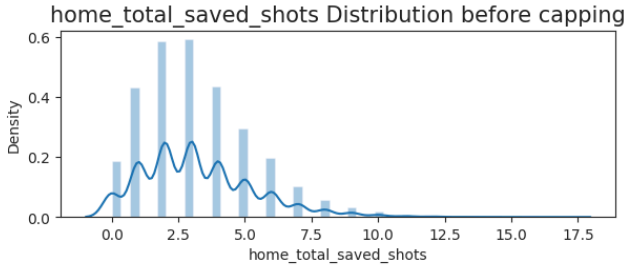

home_total_yellow_cards Distribution after capping
Density
home_total_yellow_cards

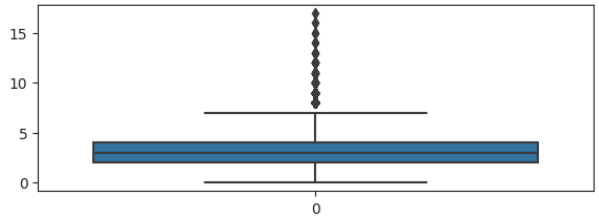

home_total_red_cards Distribution before capping
Density
home_total_red_cards

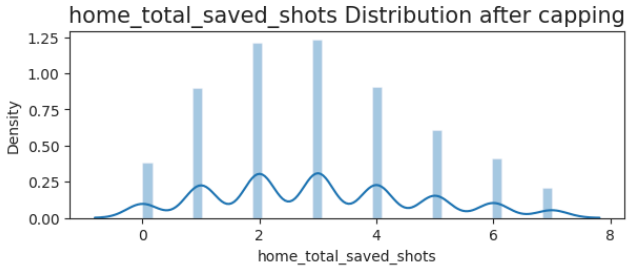

home_total_red_cards Distribution after capping
Density
home_total_red_cards

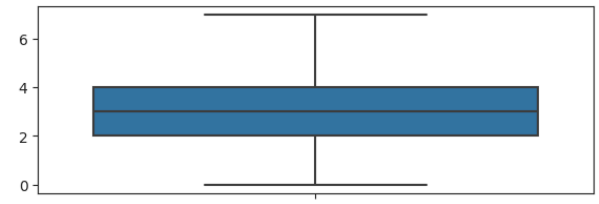

home_total_blocked_shots Distribution before capping
Density
home_total_blocked_shots

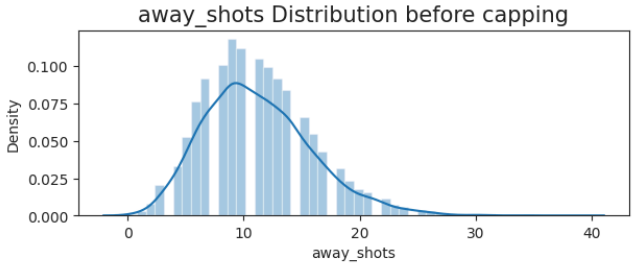

home_total_blocked_shots Distribution after capping
Density
home_total_blocked_shots

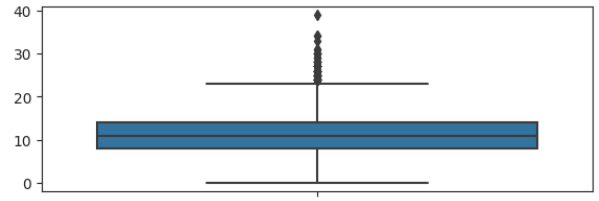

home_total_saved_shots Distribution before capping
Density
home_total_saved_shots

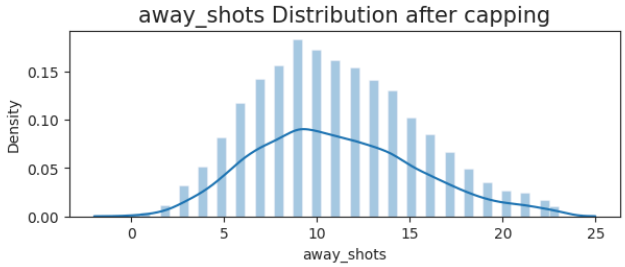


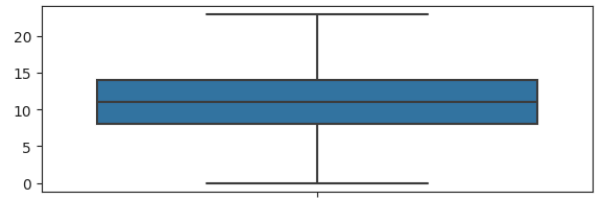

home_total_saved_shots Distribution after capping
Density
home_total_saved_shots

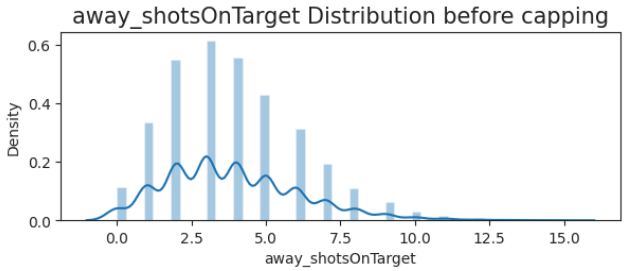


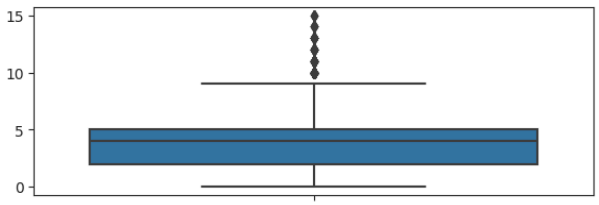

away_shots Distribution before capping
Density
away_shots

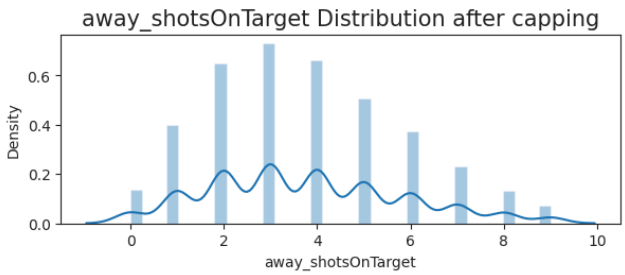


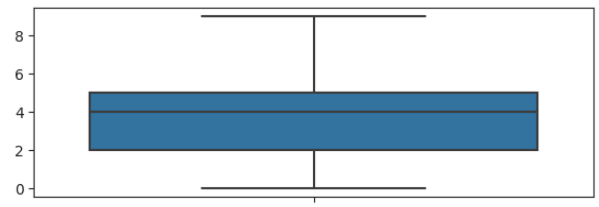

away_shots Distribution after capping
Density
away_shots

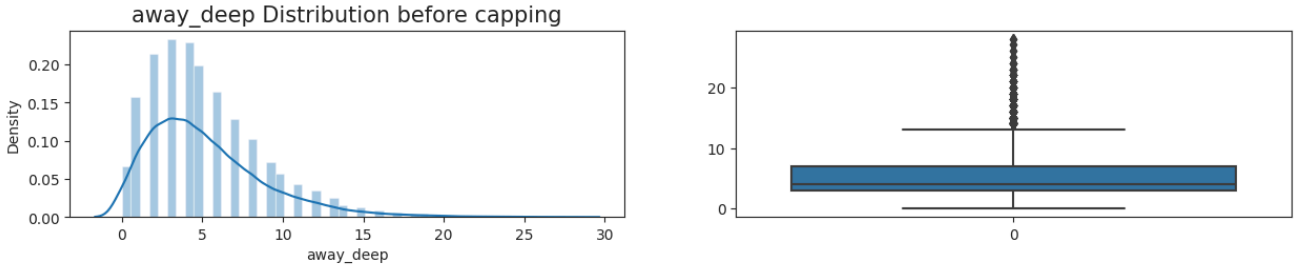


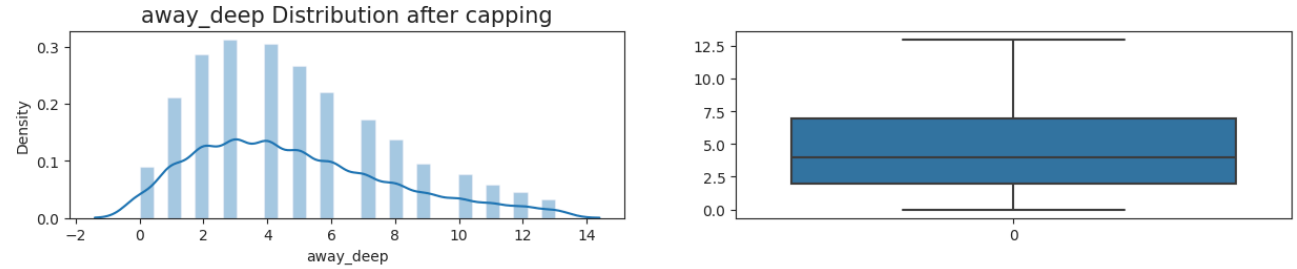

away_shotsOnTarget Distribution before capping
Density
away_shotsOnTarget

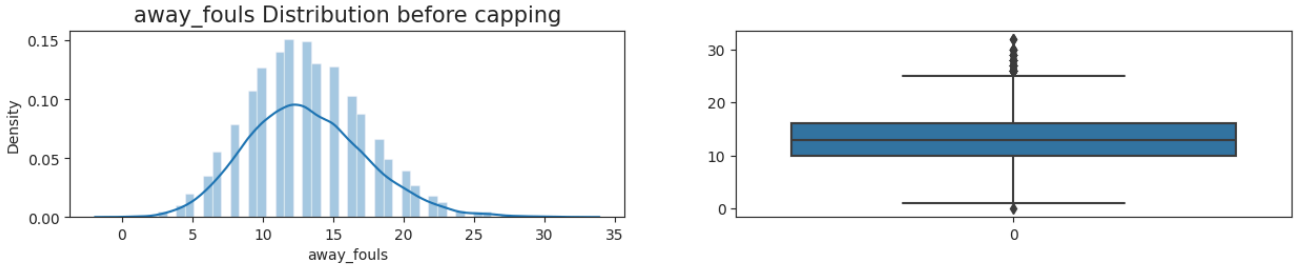


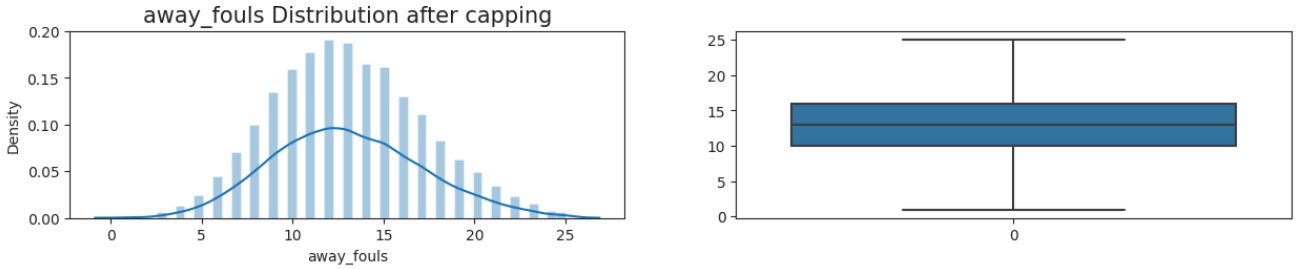

away_shotsOnTarget Distribution after capping
Density
away_shotsOnTarget

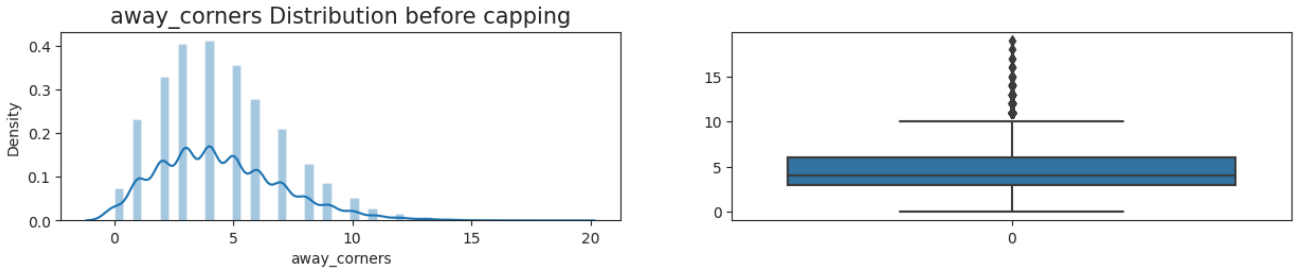


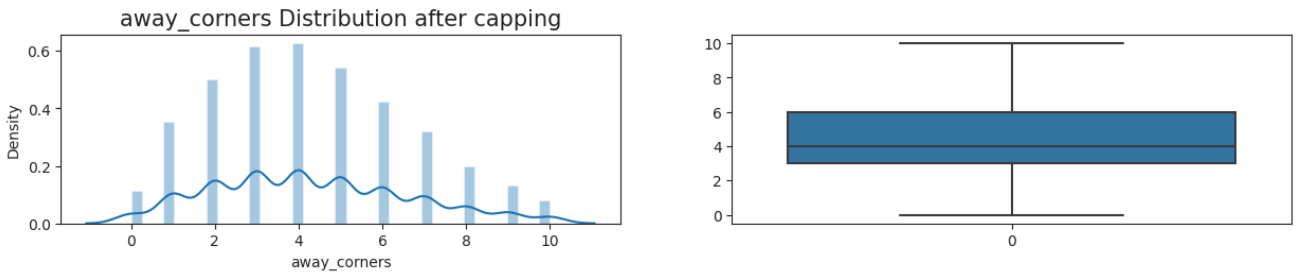

away_deep Distribution before capping
Density
away_deep

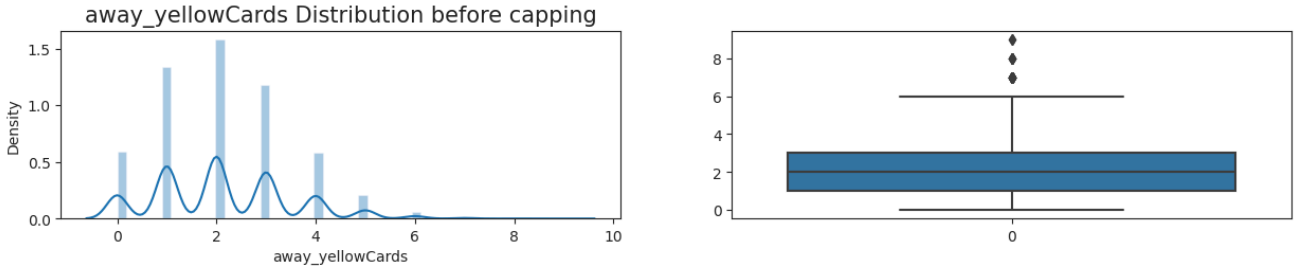

away_deep Distribution after capping
Density
away_deep

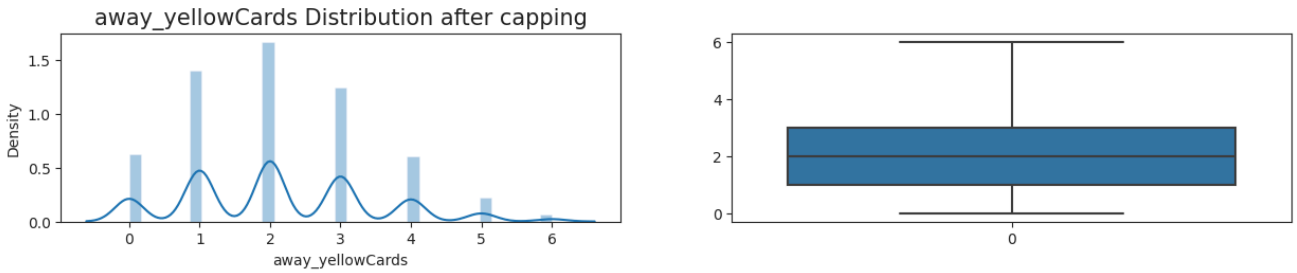

away_fouls Distribution before capping
Density
away_fouls

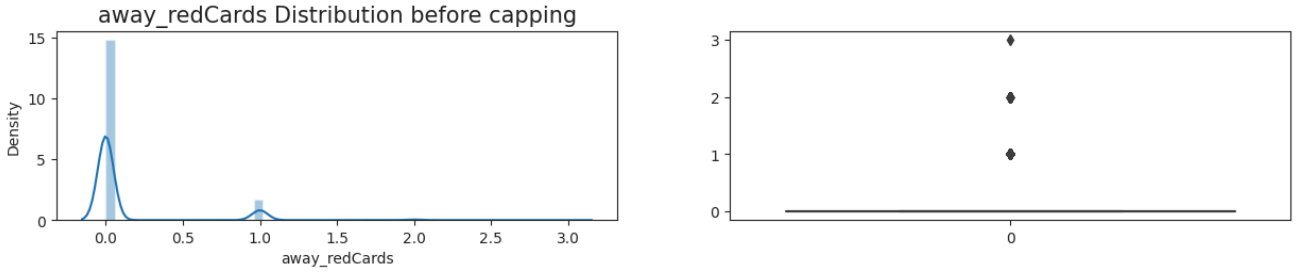

away_fouls Distribution after capping
Density
away_fouls

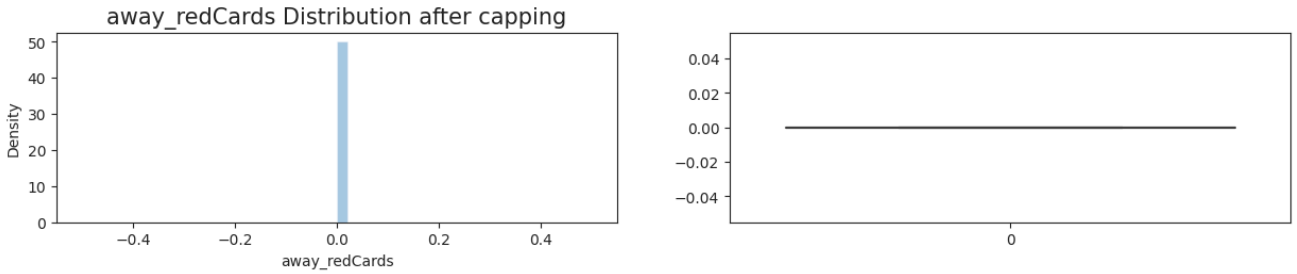

away_corners Distribution before capping
Density
away_corners

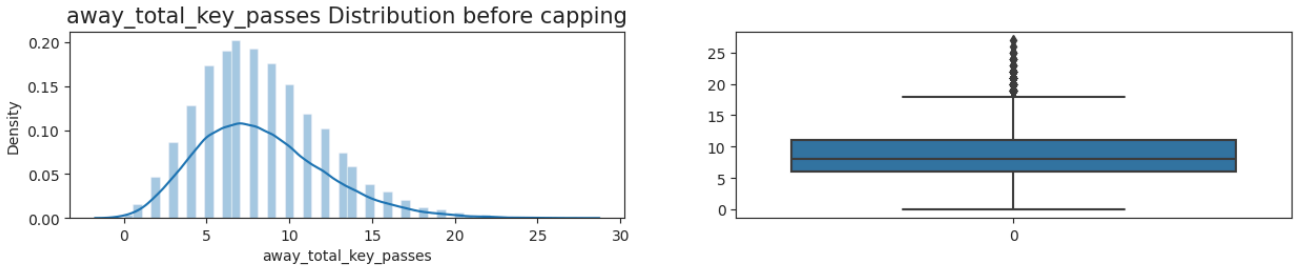

away_corners Distribution after capping
Density
away_corners

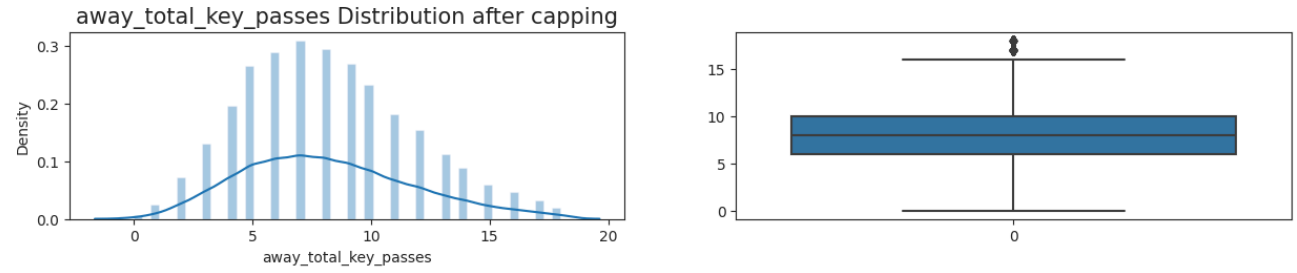

away_yellowCards Distribution before capping
Density
away_yellowCards

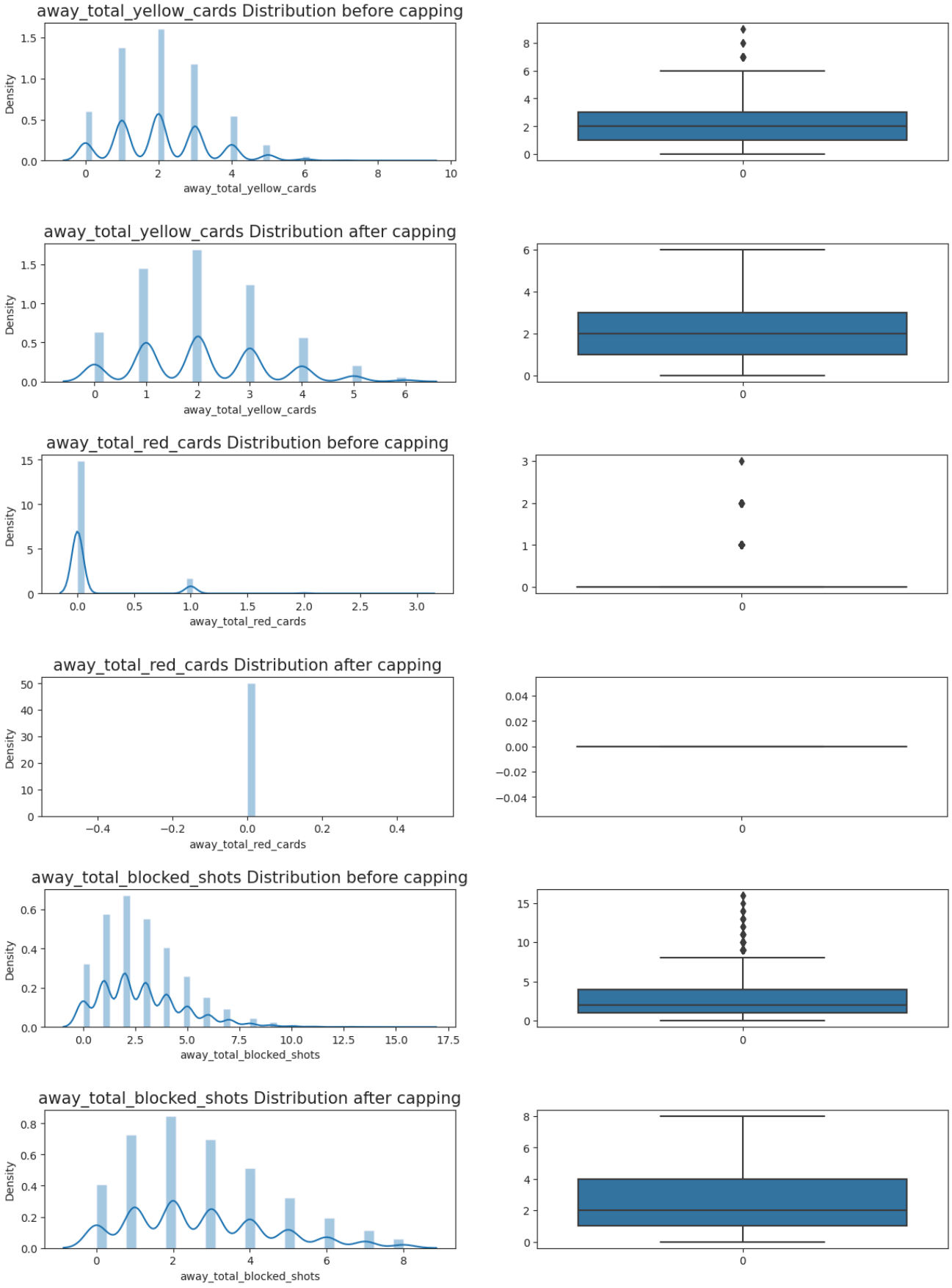

away_yellowCards Distribution after capping
Density
away_yellowCards

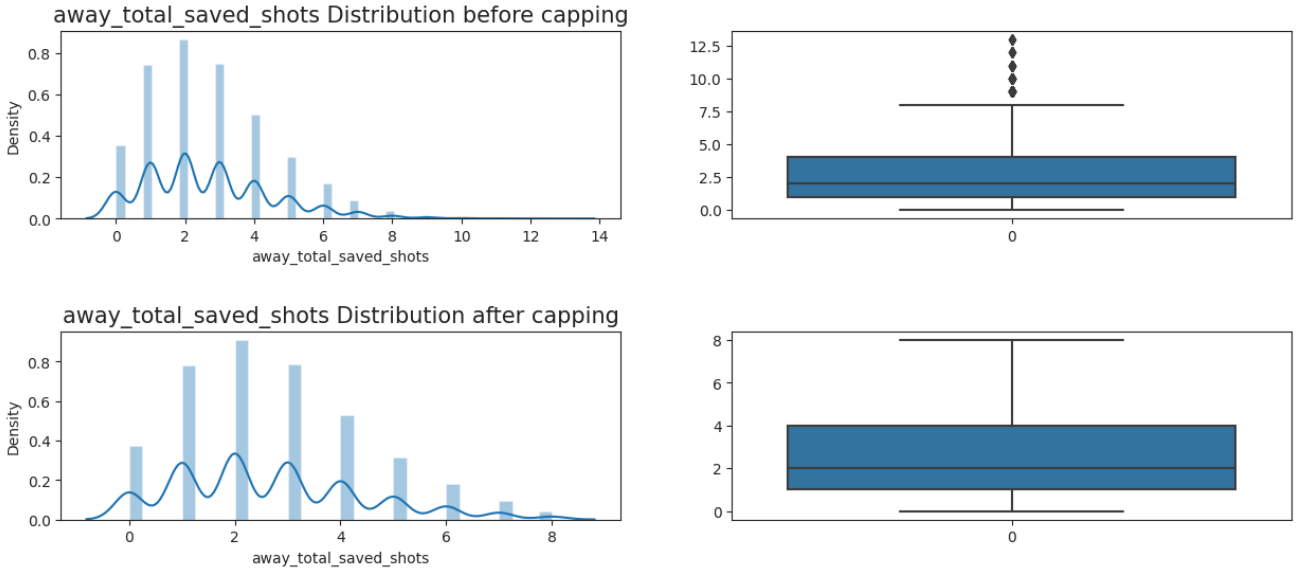

away_redCards Distribution before capping
Density
away_redCards


away_redCards Distribution after capping
Density
away_redCards


away_total_key_passes Distribution before capping
Density
away_total_key_passes


away_total_key_passes Distribution after capping
Density
away_total_key_passes

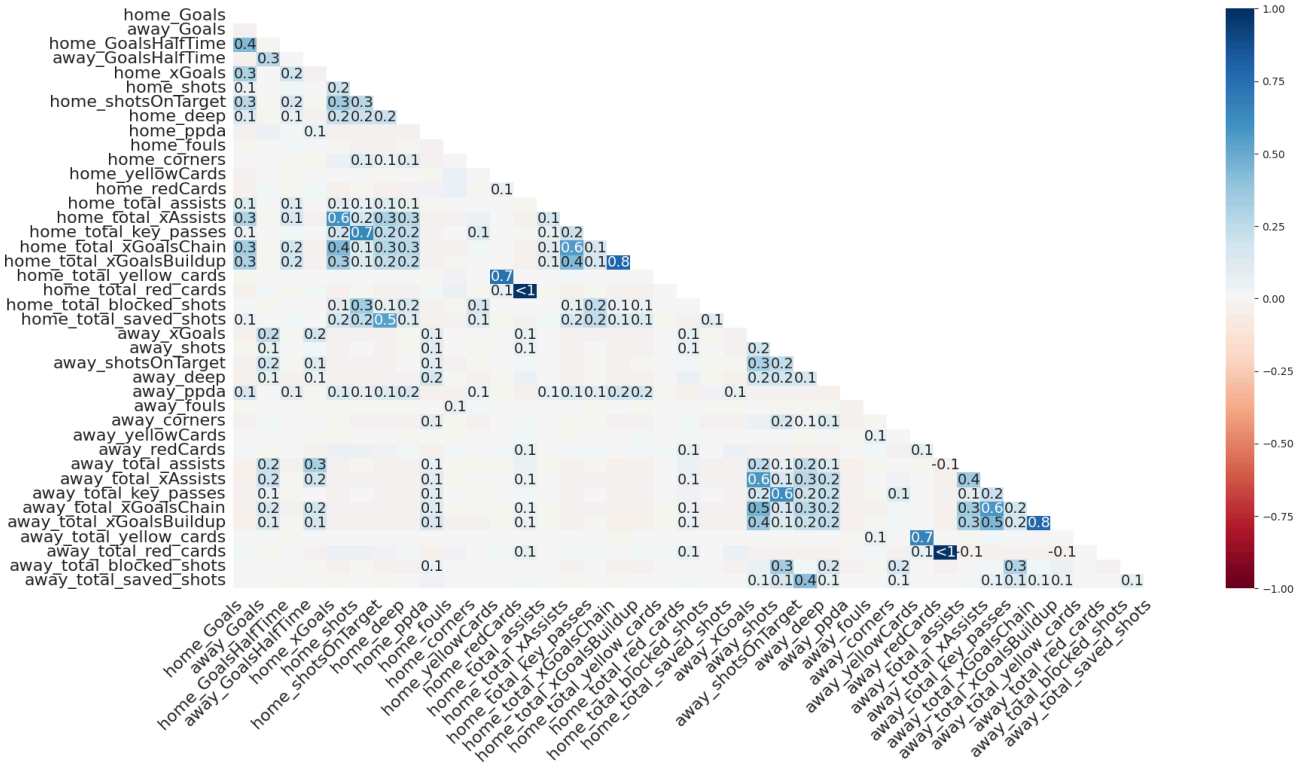

away_total_yellow_cards Distribution before capping
Density
away_total_yellow_cards
away_total_yellow_cards Distribution after capping
Density
away_total_yellow_cards
away_total_red_cards Distribution before capping
Density
away_total_red_cards
away_total_red_cards Distribution after capping
Density
away_total_red_cards
away_total_blocked_shots Distribution before capping
Density
away_total_blocked_shots
away_total_blocked_shots Distribution after capping
Density
away_total_blocked_shots



Data Cleansing

missingness correlation heatmap

<Axes: >



=== away_Goals ===

```
1    4458
0    4031
2    2522
3    1108
4     395
5     122
6      31
7       8
8       3
9       2
Name: away_Goals, dtype: int64
=== home_shotsOnTarget ===
4     2033
3     2011
5     1860
6     1477
2     1469
7     1071
1      756
8      680
9      469
10     274
0      229
11     164
12      96
13      44
14      30
15      10
16       4
17       2
18       1
Name: home_shotsOnTarget, dtype: int64
=== home_yellowCards ===
2    3631
1    3433
3    2252
0    1813
4    1038
5     348
6     128
7      30
8       7
Name: home_yellowCards, dtype: int64
=== home_redCards ===
0    11602
1     1016
2       61
3        1
Name: home_redCards, dtype: int64
=== home_total_assists ===
0    4785
1    4382
2    2288
3     858
4     264
5      83
6      14
7       5
8       1
Name: home_total_assists, dtype: int64
```

```
=== home_total_yellow_cards ===
2    3639
1    3516
3    2255
0    1830
4     997
5     318
6     105
7      16
8       4
Name: home_total_yellow_cards, dtype: int64
=== home_total_red_cards ===
0    11616
1     1007
2       56
3        1
Name: home_total_red_cards, dtype: int64
=== home_total_blocked_shots ===
3.0     2338
2.0     2318
4.0     1810
1.0     1766
5.0     1382
6.0      927
0.0      821
7.0      573
8.0      324
9.0      188
10.0     100
11.0      51
12.0      39
13.0      24
14.0       6
16.0       3
15.0       3
17.0       2
19.0       1
18.0       1
Name: home_total_blocked_shots, dtype: int64
=== home_total_saved_shots ===
3.0     2550
2.0     2515
4.0     1877
1.0     1857
5.0     1264
6.0      848
0.0      798
7.0      436
8.0      250
9.0      145
10.0      73
11.0      32
12.0      18
13.0       5
14.0       3
15.0       3
16.0       2
17.0       1
Name: home_total_saved_shots, dtype: int64
=== away_shotsOnTarget ===
```

```
3     2333
4     2114
2     2081
5     1627
1     1277
6     1191
7      738
0      431
8      423
9      232
10     115
11      64
12      27
13      15
14       8
15       4
Name: away_shotsOnTarget, dtype: int64
=== away_corners ===
4     1987
3     1952
5     1723
2     1595
6     1345
1     1121
7     1016
8      633
9      417
0      356
10     252
11     132
12      74
13      42
14      22
15       6
16       3
17       2
19       1
18       1
Name: away_corners, dtype: int64
=== away_yellowCards ===
2    3605
1    3045
3    2691
0    1356
4    1319
5     480
6     141
7      35
8       6
9       2
Name: away_yellowCards, dtype: int64
=== away_redCards ===
0    11284
1     1317
2       78
3        1
Name: away_redCards, dtype: int64
=== away_total_yellow_cards ===
```

```
2    3662
1    3144
3    2688
0    1376
4    1235
5     436
6     113
7      23
8       2
9       1
Name: away_total_yellow_cards, dtype: int64
=== away_total_red_cards ===
0    11298
1     1306
2       75
3        1
Name: away_total_red_cards, dtype: int64
=== away_total_blocked_shots ===
2.0     2722
1.0     2343
3.0     2242
4.0     1649
0.0     1311
5.0     1046
6.0      618
7.0      371
8.0      183
9.0       97
10.0      45
11.0      22
13.0      10
14.0       5
12.0       5
16.0       2
15.0       1
Name: away_total_blocked_shots, dtype: int64
=== away_total_saved_shots ===
2.0     2855
3.0     2471
1.0     2453
4.0     1655
0.0     1172
5.0      988
6.0      566
7.0      293
8.0      124
9.0       59
10.0      20
11.0       9
12.0       4
13.0       3
Name: away_total_saved_shots, dtype: int64
features with few unique values that can be treated as categorical:
 ['away_Goals', 'home_shotsOnTarget', 'home_yellowCards', 'home_redCards', 'home_t
otal_assists', 'home_total_yellow_cards', 'home_total_red_cards', 'home_total_bloc
ked_shots', 'home_total_saved_shots', 'away_shotsOnTarget', 'away_corners', 'away_
yellowCards', 'away_redCards', 'away_total_yellow_cards', 'away_total_red_cards',
'away_total_blocked_shots', 'away_total_saved_shots']
```

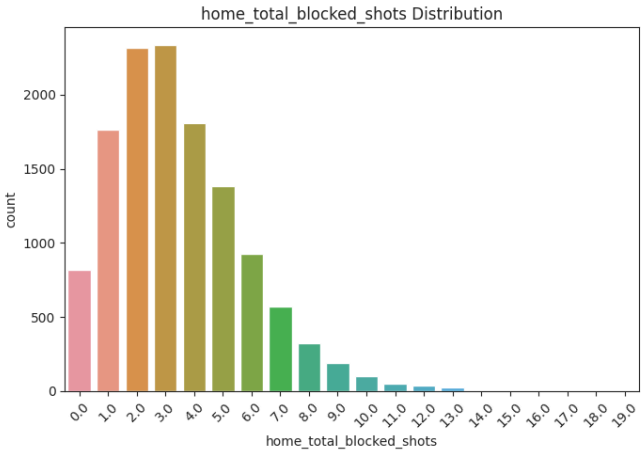
all the features here are related to the cards recieved, it was already made as a categorical feature.

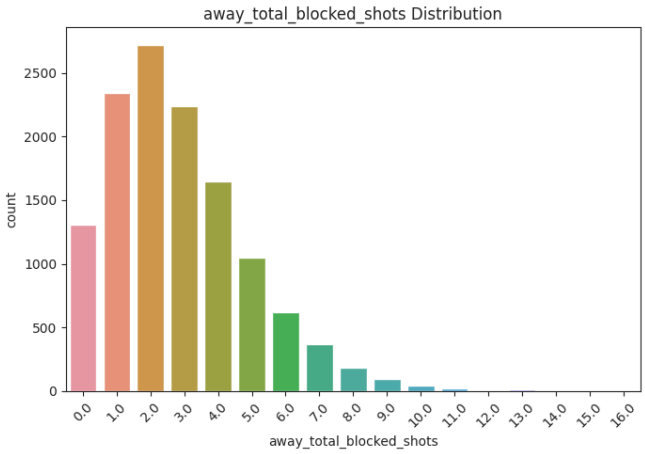
It is more likely to transform those features to categoricals.

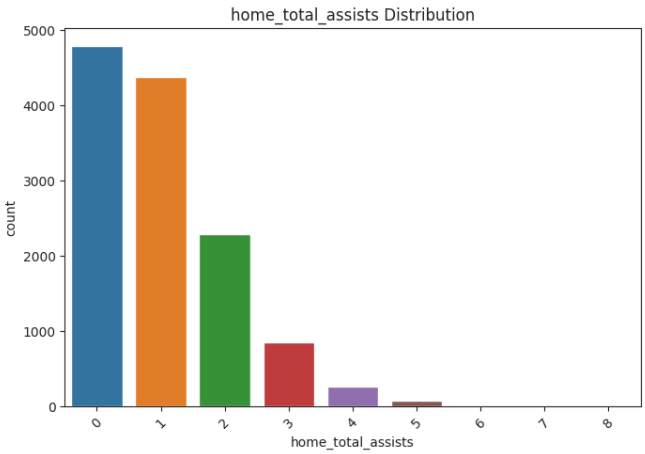
```
gameID
leagueID
season
date
homeTeamID
awayTeamID
home_Goals
away_Goals
home_GoalsHalfTime
away_GoalsHalfTime
home_xGoals
home_shots
home_shotsOnTarget
home_deep
home_ppda
home_fouls
home_corners
home_yellowCards
home_redCards
home_total_assists
home_total_xAssists
home_total_key_passes
home_total_xGoalsChain
home_total_xGoalsBuildup
home_total_yellow_cards
home_total_red_cards
home_total_blocked_shots
home_total_saved_shots
away_xGoals
away_shots
away_shotsOnTarget
away_deep
away_ppda
away_fouls
away_corners
away_yellowCards
away_redCards
away_total_assists
away_total_xAssists
away_total_key_passes
away_total_xGoalsChain
away_total_xGoalsBuildup
away_total_yellow_cards
away_total_red_cards
away_total_blocked_shots
away_total_saved_shots
gameresult
home_redCards_binary
away_redCards_binary
home_yellowCards_cat
away_yellowCards_cat
```

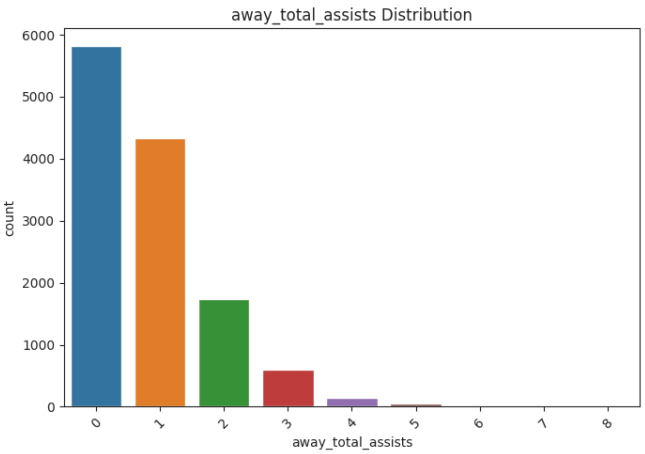
```
['away_Goals',
 'home_shotsOnTarget',
 'home_total_assists',
 'home_total_blocked_shots',
 'home_total_saved_shots',
 'away_shotsOnTarget',
 'away_corners',
 'away_total_blocked_shots',
 'away_total_saved_shots']
```

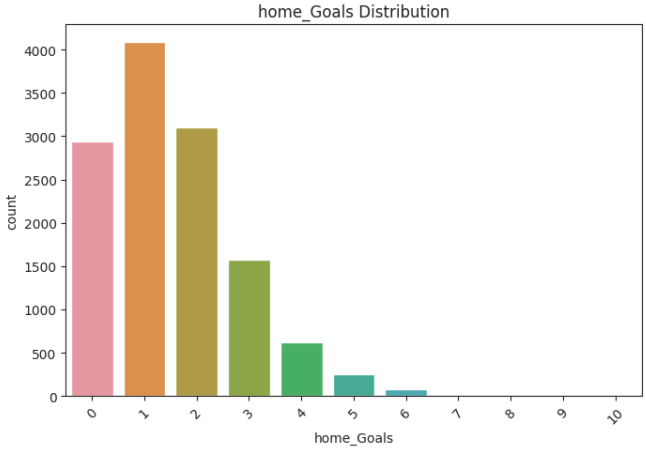
```
['away_shotsOnTarget',
 'away_total_assists',
 'away_total_blocked_shots',
 'home_total_blocked_shots',
 'away_corners',
 'home_total_assists',
 'away_total_saved_shots',
 'away_Goals',
 'home_Goals',
 'home_total_saved_shots',
 'home_shotsOnTarget',
 'home_corners']
```

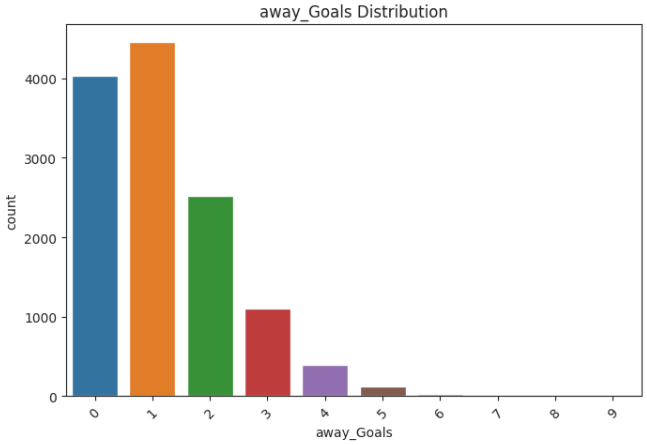


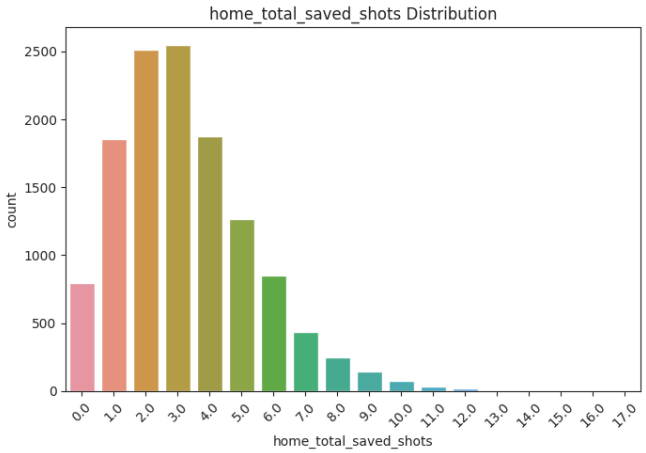


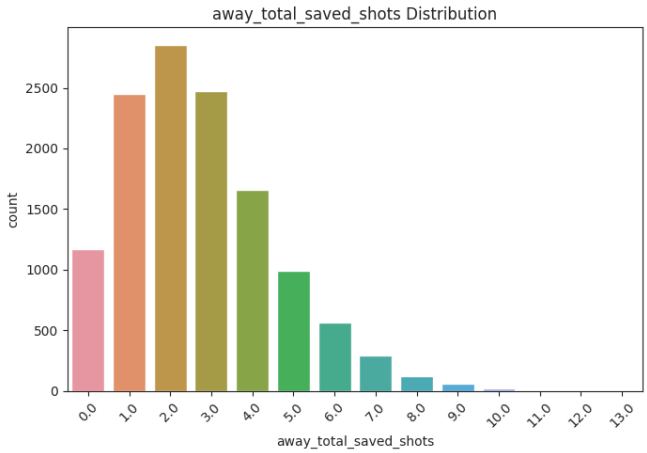


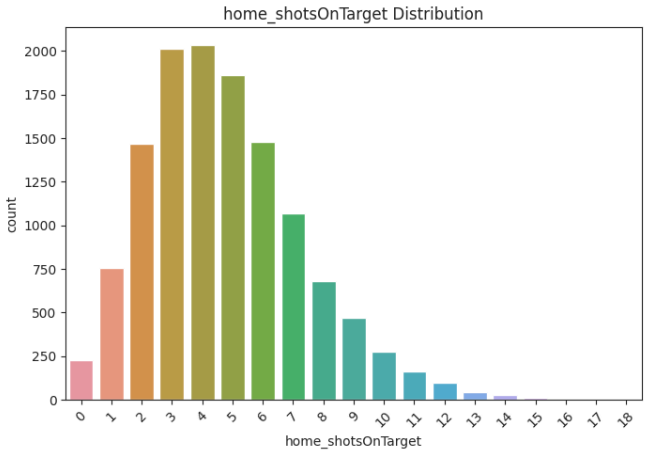


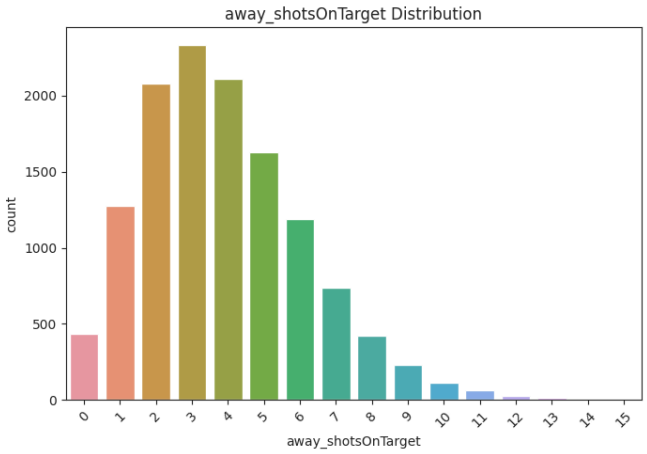


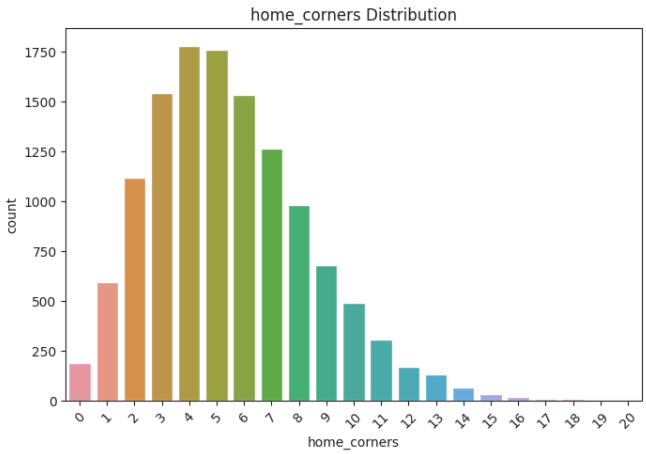


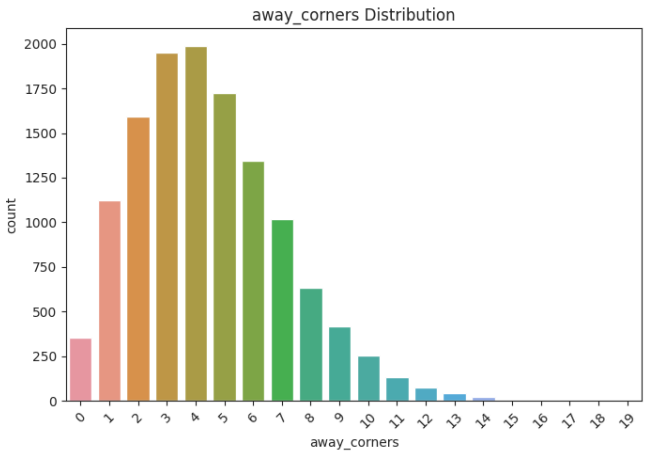












lets add categorical features coresponding to each feature in here, this is the way to deal with outliers in this case:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12680 entries, 0 to 12679
Data columns (total 63 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   gameID                        12680 non-null  int64
 1   leagueID                      12680 non-null  int64
 2   season                        12680 non-null  int64
 3   date                          12680 non-null  datetime64[ns]
 4   homeTeamID                    12680 non-null  int64
 5   awayTeamID                    12680 non-null  int64
 6   home_Goals                    12680 non-null  int64
 7   away_Goals                    12680 non-null  int64
 8   home_GoalsHalfTime            12680 non-null  int64
 9   away_GoalsHalfTime            12680 non-null  int64
 10  home_xGoals                   12680 non-null  float64
 11  home_shots                    12680 non-null  int64
 12  home_shotsOnTarget            12680 non-null  int64
 13  home_deep                     12680 non-null  int64
 14  home_ppda                     12680 non-null  float64
 15  home_fouls                    12680 non-null  int64
 16  home_corners                  12680 non-null  int64
 17  home_yellowCards              12680 non-null  int64
 18  home_redCards                 12680 non-null  int64
 19  home_total_assists            12680 non-null  int64
 20  home_total_xAssists           12680 non-null  float64
 21  home_total_key_passes         12680 non-null  int64
 22  home_total_xGoalsChain        12680 non-null  float64
 23  home_total_xGoalsBuildup      12680 non-null  float64
 24  home_total_yellow_cards       12680 non-null  int64
 25  home_total_red_cards          12680 non-null  int64
 26  home_total_blocked_shots      12677 non-null  float64
 27  home_total_saved_shots        12677 non-null  float64
 28  away_xGoals                   12680 non-null  float64
 29  away_shots                    12680 non-null  int64
 30  away_shotsOnTarget            12680 non-null  int64
 31  away_deep                     12680 non-null  int64
 32  away_ppda                     12680 non-null  float64
 33  away_fouls                    12680 non-null  int64
 34  away_corners                  12680 non-null  int64
 35  away_yellowCards              12680 non-null  int64
 36  away_redCards                 12680 non-null  int64
 37  away_total_assists            12680 non-null  int64
 38  away_total_xAssists           12680 non-null  float64
 39  away_total_key_passes         12680 non-null  int64
 40  away_total_xGoalsChain        12680 non-null  float64
 41  away_total_xGoalsBuildup      12680 non-null  float64
 42  away_total_yellow_cards       12680 non-null  int64
 43  away_total_red_cards          12680 non-null  int64
 44  away_total_blocked_shots      12672 non-null  float64
 45  away_total_saved_shots        12672 non-null  float64
 46  gameresult                    12680 non-null  int64
 47  home_redCards_binary          12680 non-null  object
 48  away_redCards_binary          12680 non-null  object
 49  home_yellowCards_cat          12680 non-null  object
 50  away_yellowCards_cat          12680 non-null  object
 51  home_shotsOnTarget_cat        12680 non-null  category
 52  away_shotsOnTarget_cat        12680 non-null  category
 53  home_total_assists_cat        12680 non-null  category
```

```
 54  away_total_assists_cat        12680 non-null  category
 55  home_corners_cat              12680 non-null  category
 56  away_corners_cat              12680 non-null  category
 57  home_Goals_cat                12680 non-null  category
 58  away_Goals_cat                12680 non-null  category
 59  home_total_blocked_shots_cat  12677 non-null  category
 60  away_total_blocked_shots_cat  12672 non-null  category
 61  home_total_saved_shots_cat    12677 non-null  category
 62  away_total_saved_shots_cat    12672 non-null  category
dtypes: category(12), datetime64[ns](1), float64(14), int64(32), object(4)
memory usage: 5.7+ MB
```

noticed that the cards categories are dtype 'object', lets change it:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12680 entries, 0 to 12679
Data columns (total 63 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   gameID                        12680 non-null  int64
 1   leagueID                      12680 non-null  int64
 2   season                        12680 non-null  int64
 3   date                          12680 non-null  datetime64[ns]
 4   homeTeamID                    12680 non-null  int64
 5   awayTeamID                    12680 non-null  int64
 6   home_Goals                    12680 non-null  int64
 7   away_Goals                    12680 non-null  int64
 8   home_GoalsHalfTime            12680 non-null  int64
 9   away_GoalsHalfTime            12680 non-null  int64
 10  home_xGoals                   12680 non-null  float64
 11  home_shots                    12680 non-null  int64
 12  home_shotsOnTarget            12680 non-null  int64
 13  home_deep                     12680 non-null  int64
 14  home_ppda                     12680 non-null  float64
 15  home_fouls                    12680 non-null  int64
 16  home_corners                  12680 non-null  int64
 17  home_yellowCards              12680 non-null  int64
 18  home_redCards                 12680 non-null  int64
 19  home_total_assists            12680 non-null  int64
 20  home_total_xAssists           12680 non-null  float64
 21  home_total_key_passes         12680 non-null  int64
 22  home_total_xGoalsChain        12680 non-null  float64
 23  home_total_xGoalsBuildup      12680 non-null  float64
 24  home_total_yellow_cards       12680 non-null  int64
 25  home_total_red_cards          12680 non-null  int64
 26  home_total_blocked_shots      12677 non-null  float64
 27  home_total_saved_shots        12677 non-null  float64
 28  away_xGoals                   12680 non-null  float64
 29  away_shots                    12680 non-null  int64
 30  away_shotsOnTarget            12680 non-null  int64
 31  away_deep                     12680 non-null  int64
 32  away_ppda                     12680 non-null  float64
 33  away_fouls                    12680 non-null  int64
 34  away_corners                  12680 non-null  int64
 35  away_yellowCards              12680 non-null  int64
 36  away_redCards                 12680 non-null  int64
 37  away_total_assists            12680 non-null  int64
 38  away_total_xAssists           12680 non-null  float64
 39  away_total_key_passes         12680 non-null  int64
 40  away_total_xGoalsChain        12680 non-null  float64
 41  away_total_xGoalsBuildup      12680 non-null  float64
 42  away_total_yellow_cards       12680 non-null  int64
 43  away_total_red_cards          12680 non-null  int64
 44  away_total_blocked_shots      12672 non-null  float64
 45  away_total_saved_shots        12672 non-null  float64
 46  gameresult                    12680 non-null  int64
 47  home_redCards_binary          12680 non-null  bool
 48  away_redCards_binary          12680 non-null  bool
 49  home_yellowCards_cat          12680 non-null  category
 50  away_yellowCards_cat          12680 non-null  category
 51  home_shotsOnTarget_cat        12680 non-null  category
 52  away_shotsOnTarget_cat        12680 non-null  category
 53  home_total_assists_cat        12680 non-null  category
```

```
 54  away_total_assists_cat        12680 non-null  category
 55  home_corners_cat              12680 non-null  category
 56  away_corners_cat              12680 non-null  category
 57  home_Goals_cat                12680 non-null  category
 58  away_Goals_cat                12680 non-null  category
 59  home_total_blocked_shots_cat  12677 non-null  category
 60  away_total_blocked_shots_cat  12672 non-null  category
 61  home_total_saved_shots_cat    12677 non-null  category
 62  away_total_saved_shots_cat    12672 non-null  category
dtypes: bool(2), category(14), datetime64[ns](1), float64(14), int64(32)
memory usage: 5.3 MB

False    11602
True      1078
Name: home_redCards_binary, dtype: int64
False    11284
True      1396
Name: away_redCards_binary, dtype: int64


=== home_xGoals ===
1.768180    3
1.510590    3
1.221600    2
1.129840    2
1.432290    2
           ..
2.150800    1
0.996123    1
1.363870    1
0.900312    1
0.323960    1
Name: home_xGoals, Length: 12491, dtype: int64
=== home_shots ===
```

```
14    1019
11     999
12     979
13     970
10     883
15     851
9      807
16     803
8      654
17     647
18     580
7      501
19     449
6      376
20     365
21     316
22     253
5      227
23     208
24     154
25     123
4      117
26      75
3       66
27      59
28      54
29      37
2       21
30      19
31      16
33      12
32      10
1        9
34       6
36       4
37       3
38       2
0        2
43       1
35       1
39       1
47       1
Name: home_shots, dtype: int64
=== home_deep ===
```

```
4     1461
5     1417
6     1300
3     1298
7     1135
2     1107
8      884
9      748
1      677
10     538
11     435
12     342
13     271
0      230
14     203
15     135
16     120
17      87
18      65
19      62
20      42
23      28
21      27
22      22
24      11
25       9
27       6
26       6
28       4
32       2
33       1
29       1
34       1
35       1
30       1
31       1
37       1
42       1
Name: home_deep, dtype: int64
=== home_fouls ===
```

```
12    1250
13    1236
11    1183
14    1095
10    1078
15     950
9      942
16     812
8      697
17     688
7      497
18     472
19     355
6      307
20     247
5      183
21     178
22     129
4       99
23      87
24      48
3       41
25      35
26      31
2       11
28      11
27       8
29       5
1        2
0        2
33       1
Name: home_fouls, dtype: int64
=== home_corners ===
4     1778
5     1758
3     1544
6     1533
7     1265
2     1118
8      983
9      678
1      595
10     489
11     307
0      187
12     169
13     133
14      65
15      33
16      19
17      11
18       9
20       3
19       3
Name: home_corners, dtype: int64
=== home_total_xAssists ===
```

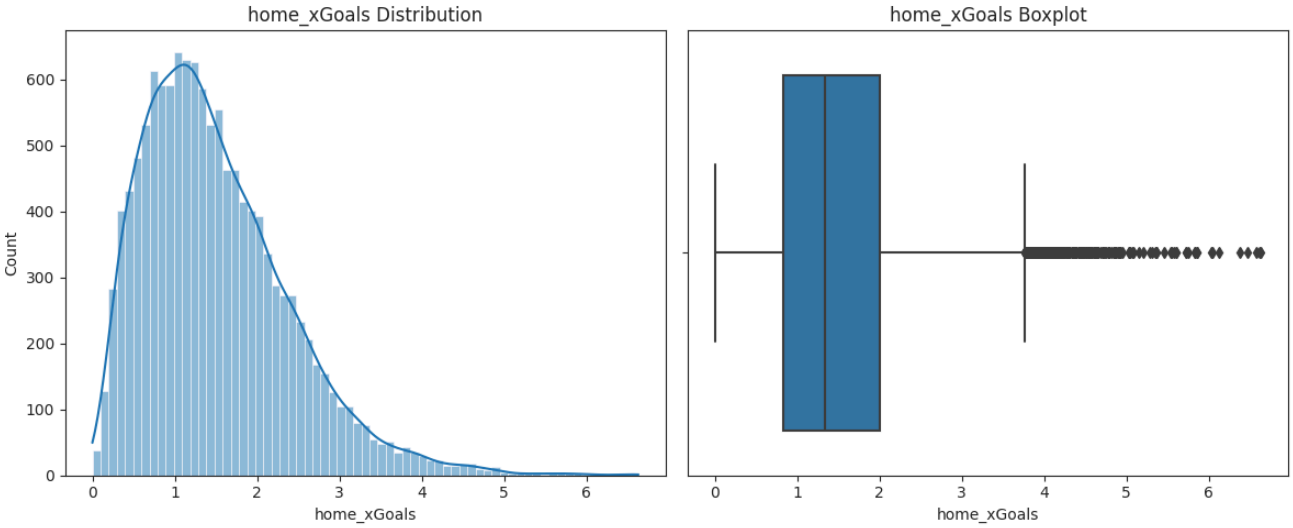
```
0.000000    9
0.284979    1
0.495288    1
1.271469    1
1.033093    1
           ..
0.578854    1
0.813590    1
0.689226    1
0.560451    1
0.074636    1
Name: home_total_xAssists, Length: 12672, dtype: int64
=== home_total_key_passes ===
9     1197
10    1194
8     1179
7     1146
11    1096
6      972
12     935
13     762
5      705
14     625
4      496
15     467
16     388
17     287
3      261
18     233
19     164
2      143
20     130
21      82
22      48
23      46
1       39
24      23
26      19
25      16
0        8
27       6
28       3
33       2
31       2
29       2
32       1
30       1
34       1
38       1
Name: home_total_key_passes, dtype: int64
=== away_shots ===
```

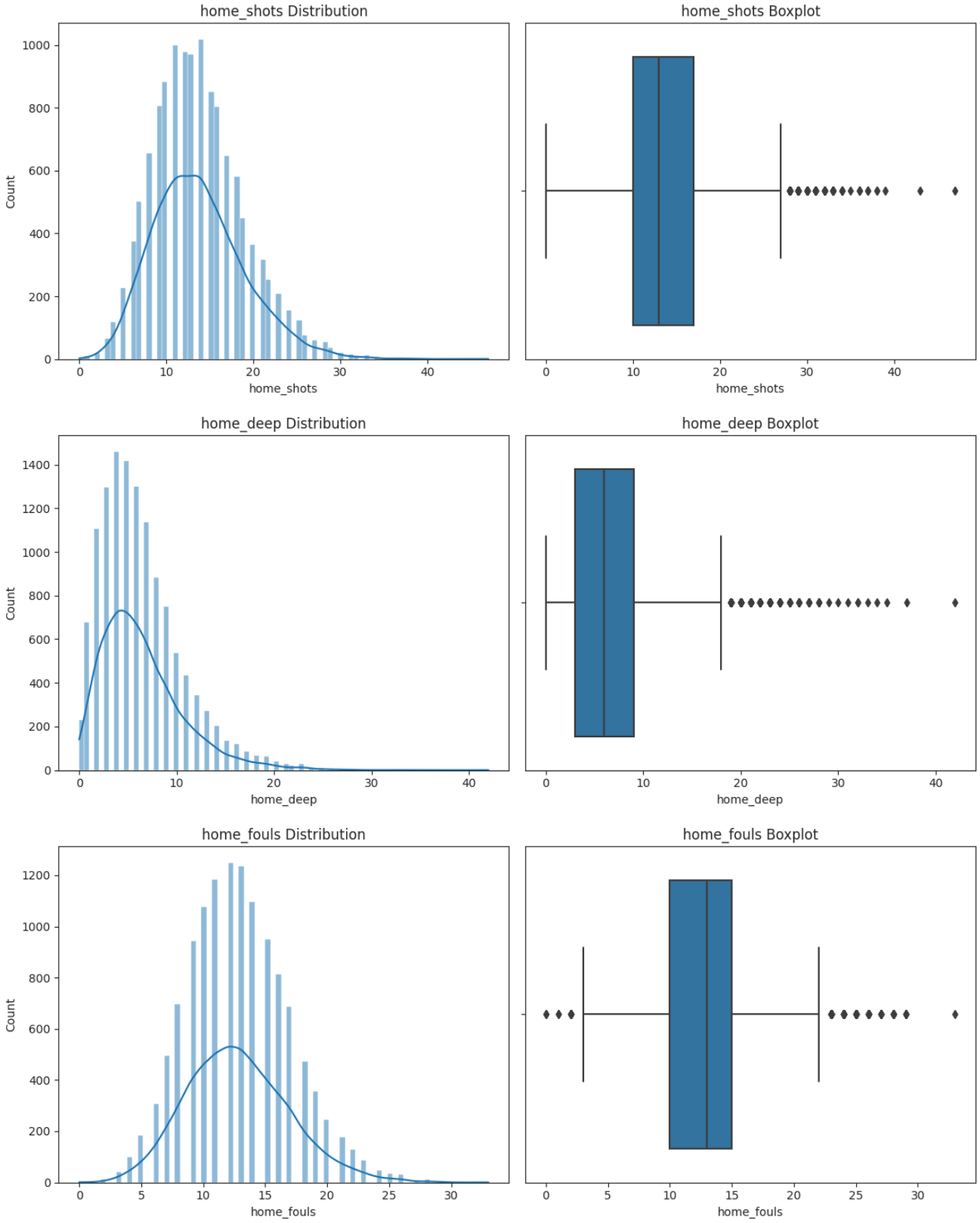
```
9     1170
10    1105
11    1037
8      998
12     983
7      907
13     905
14     831
6      754
15     651
16     540
5      523
17     425
4      328
18     316
19     228
3      203
20     172
21     154
22     111
2       75
23      68
24      55
25      39
26      30
1       27
27      15
0        8
28       7
30       7
29       2
31       2
34       2
33       1
39       1
Name: away_shots, dtype: int64
=== away_deep ===
```

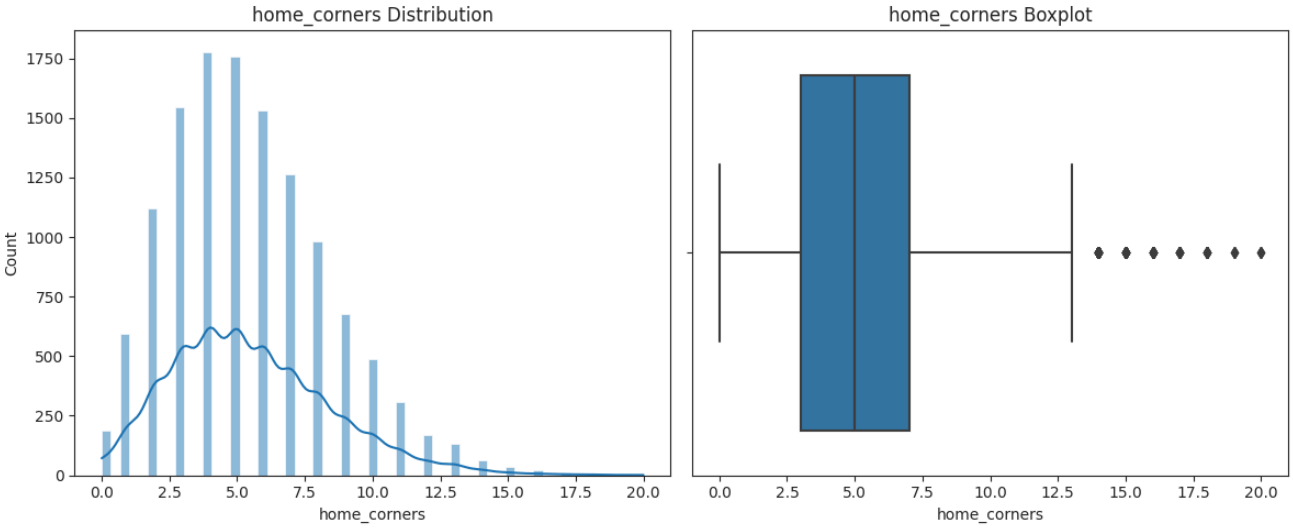
```
3     1656
4     1624
2     1520
5     1410
6     1172
1     1118
7      915
8      725
9      509
0      475
10     406
11     308
12     245
13     174
14     112
15      88
16      58
17      46
18      31
19      29
20      19
21      11
22       9
23       7
24       5
27       2
28       2
26       2
25       2
Name: away_deep, dtype: int64
=== away_fouls ===
```

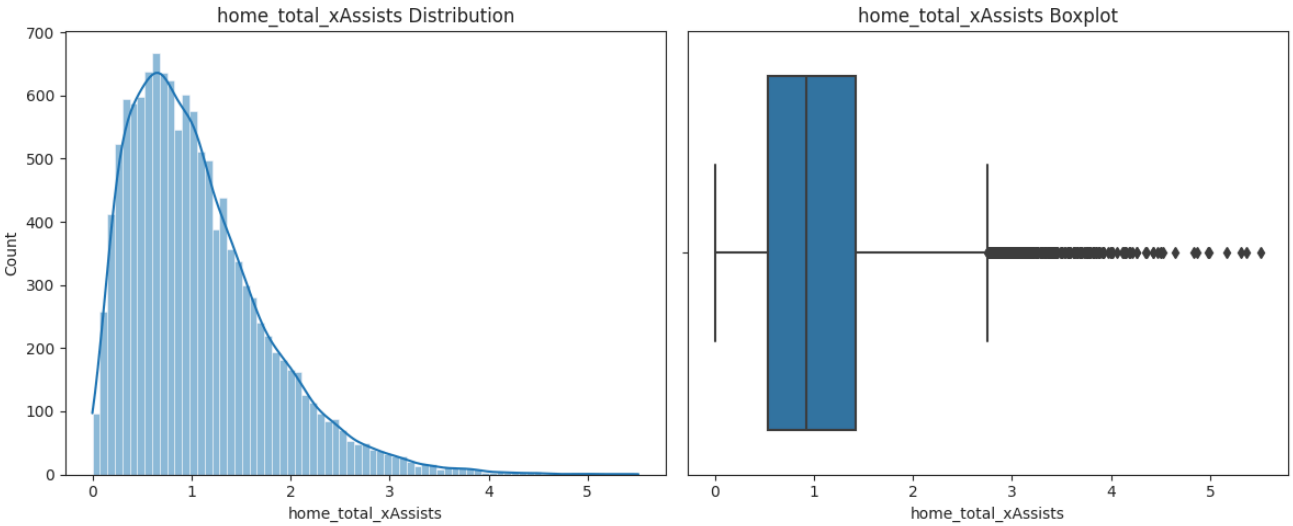
```
12    1226
13    1210
11    1142
14    1060
15    1037
10    1029
9      868
16     834
17     715
8      640
18     536
7      450
19     400
20     318
6      287
21     218
5      158
22     147
23     102
4       84
24      46
25      40
3       38
26      38
27      16
28      11
29       7
2        7
1        7
30       6
32       2
0        1
Name: away_fouls, dtype: int64
=== away_total_key_passes ===
```

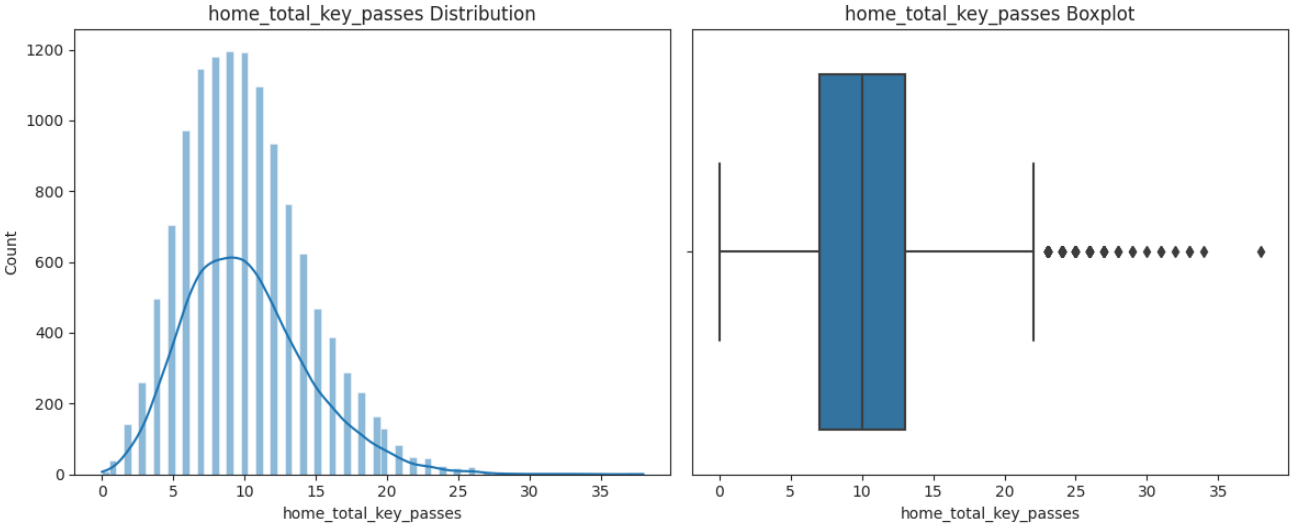
```
7     1386
8     1323
6     1302
9     1210
5     1193
10    1043
4      882
11     815
12     695
3      590
13     508
14     400
2      323
15     268
16     206
17     143
1      110
18      87
19      65
20      44
0       27
22      23
21      20
24       6
23       5
25       4
26       1
27       1
Name: away_total_key_passes, dtype: int64
features with outliers to remove:
 ['home_xGoals', 'home_shots', 'home_deep', 'home_fouls', 'home_corners', 'home_to
tal_xAssists', 'home_total_key_passes', 'away_shots', 'away_deep', 'away_fouls',
'away_total_key_passes']
```

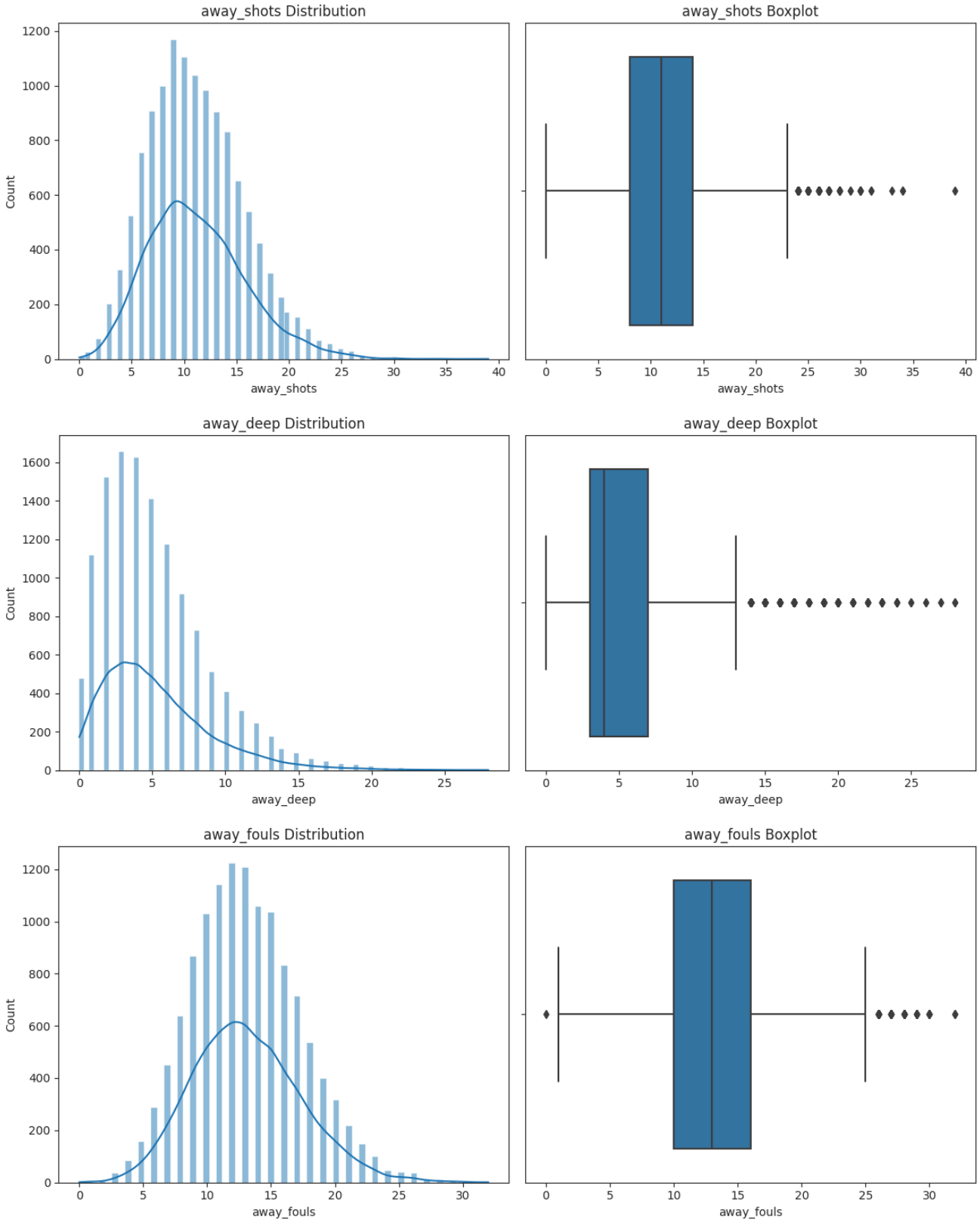


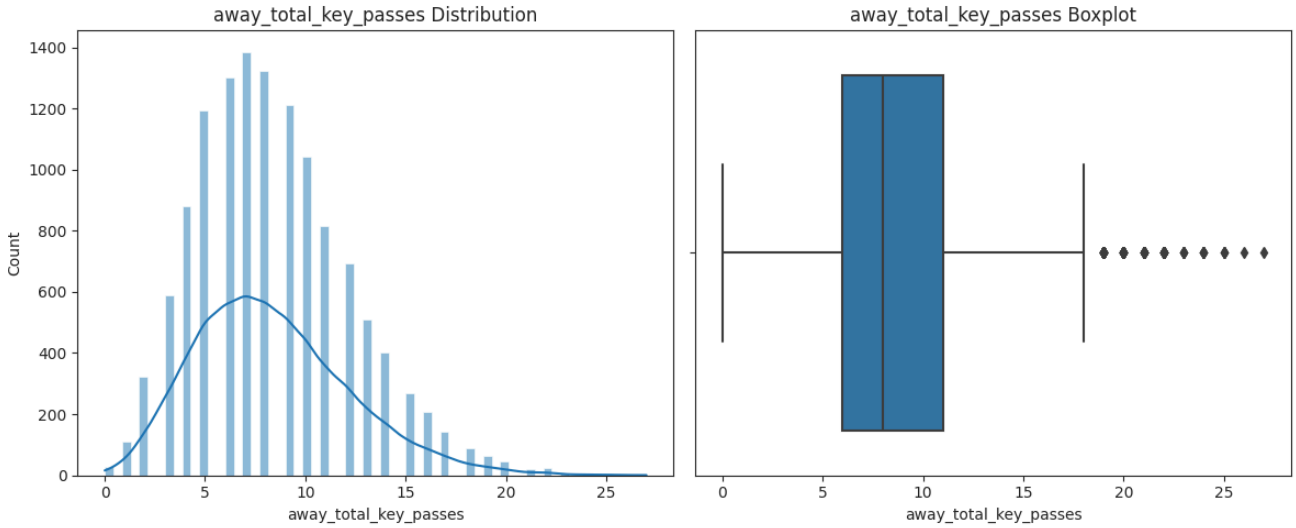

home_shots Distribution
Count
home_shots
home_shots Boxplot
home_shots
home_deep Distribution
Count
home_deep
home_deep Boxplot
home_deep
home_fouls Distribution
Count
home_fouls
home_fouls Boxplot
home_fouls


home_corners Distribution
Count
home_corners
home_corners Boxplot
home_corners


home_total_xAssists Distribution
Count
home_total_xAssists
home_total_xAssists Boxplot
home_total_xAssists


home_total_key_passes Distribution
Count
home_total_key_passes
home_total_key_passes Boxplot
home_total_key_passes


away_shots Distribution
Count
away_shots
away_shots Boxplot
away_shots
away_deep Distribution
Count
away_deep
away_deep Boxplot
away_deep
away_fouls Distribution
Count
away_fouls
away_fouls Boxplot
away_fouls




```
['home_xGoals',
 'home_shots',
 'home_deep',
 'home_fouls',
 'home_corners',
 'home_total_xAssists',
 'home_total_key_passes',
 'away_shots',
 'away_deep',
 'away_fouls',
 'away_total_key_passes']
```

	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals	ho
0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	1	0	
4	0	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	
12675	0	0	0	0	0	0	0	0	
12676	0	0	0	0	0	0	0	0	
12677	0	0	0	0	0	0	0	0	
12678	0	0	0	0	0	0	0	0	
12679	0	0	0	0	0	0	0	0	

12680 rows × 51 columns

	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
0	81	1	2015	2015-08-08 15:45:00	89	82	1.0	0.0
1	82	1	2015	2015-08-08 18:00:00	73	71	0.0	1.0
2	83	1	2015	2015-08-08 18:00:00	72	90	2.0	2.0
3	84	1	2015	2015-08-08 18:00:00	75	77	NaN	2.0
4	85	1	2015	2015-08-08 18:00:00	79	78	1.0	3.0
...	...	...	...	...	...	...	...	...
12675	16131	5	2020	2021-05-23 19:00:00	168	166	1.0	2.0
12676	16132	5	2020	2021-05-23 19:00:00	177	176	1.0	2.0
12677	16133	5	2020	2021-05-23 19:00:00	163	235	2.0	0.0
12678	16134	5	2020	2021-05-23 19:00:00	175	181	0.0	1.0
12679	16135	5	2020	2021-05-23 19:00:00	225	179	1.0	1.0

12680 rows × 51 columns

	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
0	81	1	2015	2015-08-08 15:45:00	89	82	1	0
1	82	1	2015	2015-08-08 18:00:00	73	71	0	1
2	83	1	2015	2015-08-08 18:00:00	72	90	2	2
3	84	1	2015	2015-08-08 18:00:00	75	77	4	2
4	85	1	2015	2015-08-08 18:00:00	79	78	1	3
...	...	...	...	...	...	...	...	...
12675	16131	5	2020	2021-05-23 19:00:00	168	166	1	2
12676	16132	5	2020	2021-05-23 19:00:00	177	176	1	2
12677	16133	5	2020	2021-05-23 19:00:00	163	235	2	0
12678	16134	5	2020	2021-05-23 19:00:00	175	181	0	1
12679	16135	5	2020	2021-05-23 19:00:00	225	179	1	1

12680 rows × 63 columns

```
NaN        276
1.51059      3
1.76818      3
1.49515      2
1.70719      2
          ...
1.12394      1
1.78285      1
2.93897      1
2.29008      1
0.32396      1
Name: home_xGoals, Length: 12216, dtype: int64
```

Missing Values

	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
0	81	1	2015	2015-08-08 15:45:00	89	82	1	0
1	82	1	2015	2015-08-08 18:00:00	73	71	0	1
2	83	1	2015	2015-08-08 18:00:00	72	90	2	2
3	84	1	2015	2015-08-08 18:00:00	75	77	4	2
4	85	1	2015	2015-08-08 18:00:00	79	78	1	3
...	...	...	...	...	...	...	...	...
12675	16131	5	2020	2021-05-23 19:00:00	168	166	1	2
12676	16132	5	2020	2021-05-23 19:00:00	177	176	1	2
12677	16133	5	2020	2021-05-23 19:00:00	163	235	2	0
12678	16134	5	2020	2021-05-23 19:00:00	175	181	0	1
12679	16135	5	2020	2021-05-23 19:00:00	225	179	1	1

12680 rows × 63 columns

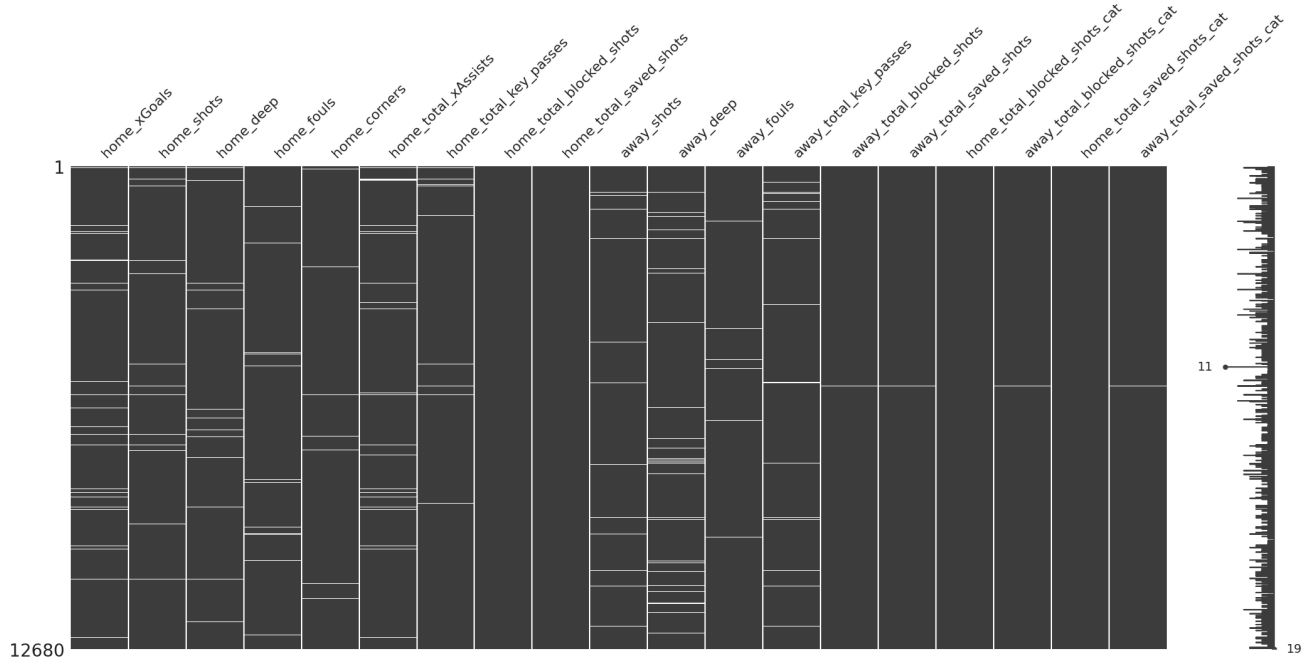
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12680 entries, 0 to 12679
Data columns (total 63 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   gameID                         12680 non-null  int64
 1   leagueID                       12680 non-null  int64
 2   season                         12680 non-null  int64
 3   date                           12680 non-null  datetime64[ns]
 4   homeTeamID                     12680 non-null  int64
 5   awayTeamID                     12680 non-null  int64
 6   home_Goals                     12680 non-null  int64
 7   away_Goals                     12680 non-null  int64
 8   home_GoalsHalfTime             12680 non-null  int64
 9   away_GoalsHalfTime             12680 non-null  int64
 10  home_xGoals                    12404 non-null  float64
 11  home_shots                     12513 non-null  float64
 12  home_shotsOnTarget             12680 non-null  int64
 13  home_deep                      12453 non-null  float64
 14  home_ppda                      12680 non-null  float64
 15  home_fouls                     12439 non-null  float64
 16  home_corners                   12537 non-null  float64
 17  home_yellowCards               12680 non-null  int64
 18  home_redCards                  12680 non-null  int64
 19  home_total_assists             12680 non-null  int64
 20  home_total_xAssists            12334 non-null  float64
 21  home_total_key_passes          12557 non-null  float64
 22  home_total_xGoalsChain         12680 non-null  float64
 23  home_total_xGoalsBuildup       12680 non-null  float64
 24  home_total_yellow_cards        12680 non-null  int64
 25  home_total_red_cards           12680 non-null  int64
 26  home_total_blocked_shots       12677 non-null  float64
 27  home_total_saved_shots         12677 non-null  float64
 28  away_xGoals                    12680 non-null  float64
 29  away_shots                     12519 non-null  float64
 30  away_shotsOnTarget             12680 non-null  int64
 31  away_deep                      12257 non-null  float64
 32  away_ppda                      12680 non-null  float64
 33  away_fouls                     12599 non-null  float64
 34  away_corners                   12680 non-null  int64
 35  away_yellowCards               12680 non-null  int64
 36  away_redCards                  12680 non-null  int64
 37  away_total_assists             12680 non-null  int64
 38  away_total_xAssists            12680 non-null  float64
 39  away_total_key_passes          12511 non-null  float64
 40  away_total_xGoalsChain         12680 non-null  float64
 41  away_total_xGoalsBuildup       12680 non-null  float64
 42  away_total_yellow_cards        12680 non-null  int64
 43  away_total_red_cards           12680 non-null  int64
 44  away_total_blocked_shots       12672 non-null  float64
 45  away_total_saved_shots         12672 non-null  float64
 46  gameresult                     12680 non-null  int64
 47  home_redCards_binary           12680 non-null  bool
 48  away_redCards_binary           12680 non-null  bool
 49  home_yellowCards_cat           12680 non-null  category
 50  away_yellowCards_cat           12680 non-null  category
 51  home_shotsOnTarget_cat         12680 non-null  category
 52  away_shotsOnTarget_cat         12680 non-null  category
 53  home_total_assists_cat         12680 non-null  category
```

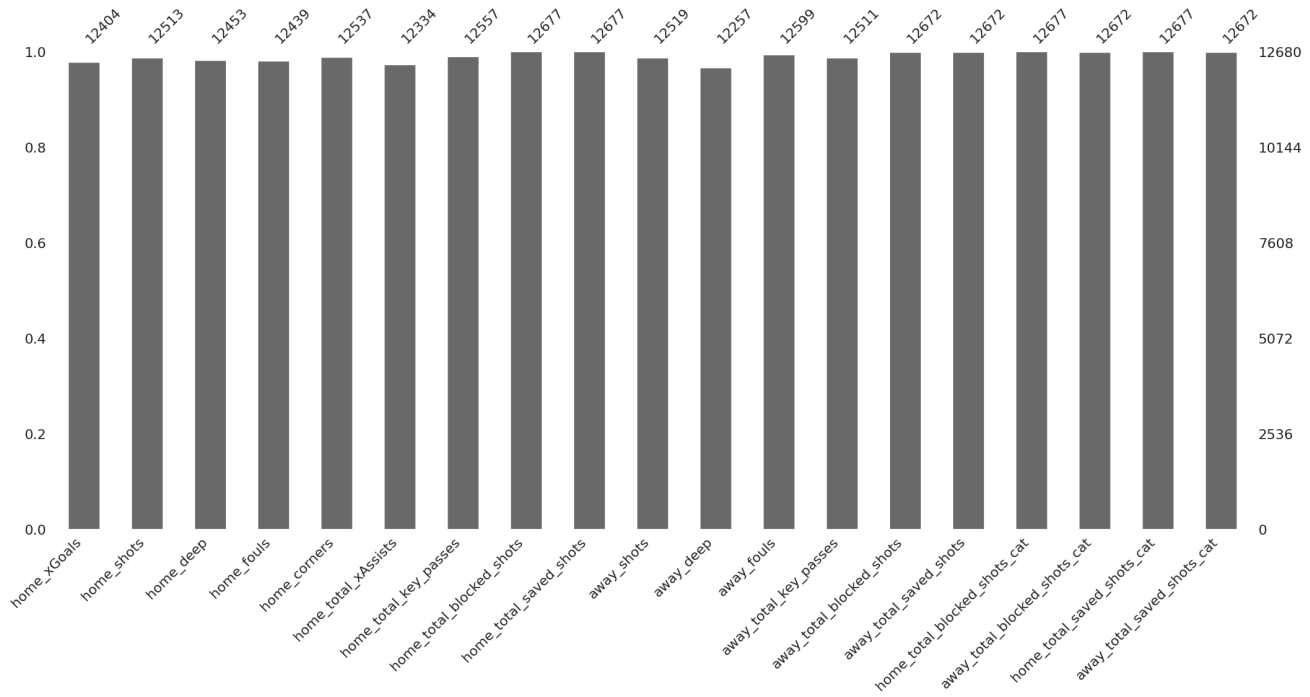
```
 54  away_total_assists_cat        12680 non-null  category
 55  home_corners_cat              12680 non-null  category
 56  away_corners_cat              12680 non-null  category
 57  home_Goals_cat                12680 non-null  category
 58  away_Goals_cat                12680 non-null  category
 59  home_total_blocked_shots_cat  12677 non-null  category
 60  away_total_blocked_shots_cat  12672 non-null  category
 61  home_total_saved_shots_cat    12677 non-null  category
 62  away_total_saved_shots_cat    12672 non-null  category
dtypes: bool(2), category(14), datetime64[ns](1), float64(23), int64(23)
memory usage: 5.3 MB
```

Now we can treat the missing data on df

	home_xGoals	home_shots	home_deep	home_fouls	home_corners	home_total_xAssists
0	0.627539	9.0	4.0	12.0	1.0	0.284979
1	0.876106	11.0	11.0	13.0	6.0	0.419975
2	0.604226	10.0	5.0	7.0	8.0	0.549139
3	2.568030	19.0	5.0	13.0	6.0	1.727543
4	1.130760	17.0	5.0	14.0	1.0	0.416638
...	...	...	...	...	...	...
12675	1.411190	15.0	17.0	8.0	9.0	0.971853
12676	1.198190	10.0	3.0	11.0	5.0	0.855524
12677	1.332690	12.0	10.0	11.0	4.0	1.151649
12678	1.460500	19.0	6.0	13.0	9.0	1.265829
12679	0.323960	6.0	1.0	17.0	2.0	0.074636

12680 rows × 19 columns





Updated missing values count and frequency after outliers removal:

	Missing Values	% of Total Values
away_deep	423	3.3
home_total_xAssists	346	2.7
home_xGoals	276	2.2
home_fouls	241	1.9
home_deep	227	1.8
away_total_key_passes	169	1.3
home_shots	167	1.3
away_shots	161	1.3
home_corners	143	1.1
home_total_key_passes	123	1.0
away_fouls	81	0.6
away_total_blocked_shots	8	0.1
away_total_saved_shots	8	0.1
away_total_blocked_shots_cat	8	0.1
away_total_saved_shots_cat	8	0.1
home_total_saved_shots	3	0.0
home_total_blocked_shots	3	0.0
home_total_blocked_shots_cat	3	0.0
home_total_saved_shots_cat	3	0.0

```
Missing Values       2401.0
% of Total Values      18.9
dtype: float64
```

Creating a dataframe with missing values as 1 and existing values as 0:

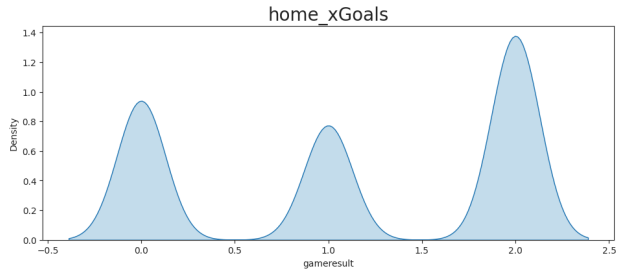
	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals	ho
0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	
12675	0	0	0	0	0	0	0	0	
12676	0	0	0	0	0	0	0	0	
12677	0	0	0	0	0	0	0	0	
12678	0	0	0	0	0	0	0	0	
12679	0	0	0	0	0	0	0	0	

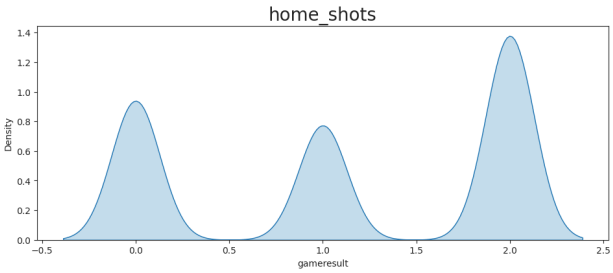
12680 rows × 63 columns

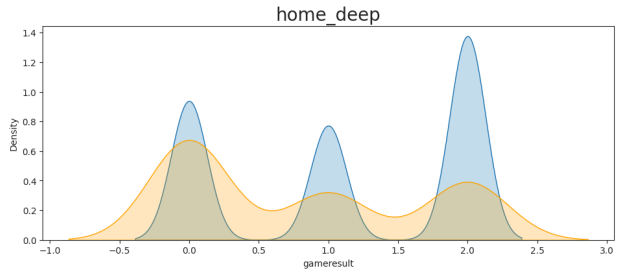
Creating df including numeric features of the later data - manipulated_data

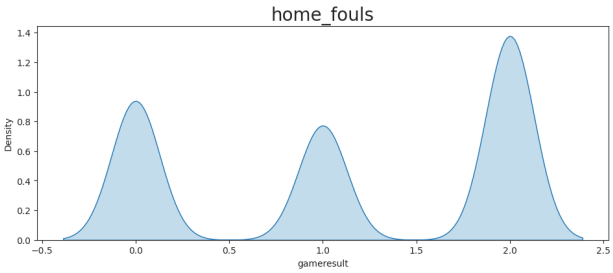
```
Index(['gameID', 'leagueID', 'season', 'homeTeamID', 'awayTeamID',
       'home_Goals', 'away_Goals', 'home_GoalsHalfTime', 'away_GoalsHalfTime',
       'home_xGoals', 'home_shots', 'home_shotsOnTarget', 'home_deep',
       'home_ppda', 'home_fouls', 'home_corners', 'home_yellowCards',
       'home_redCards', 'home_total_assists', 'home_total_xAssists',
       'home_total_key_passes', 'home_total_xGoalsChain',
       'home_total_xGoalsBuildup', 'home_total_yellow_cards',
       'home_total_red_cards', 'home_total_blocked_shots',
       'home_total_saved_shots', 'away_xGoals', 'away_shots',
       'away_shotsOnTarget', 'away_deep', 'away_ppda', 'away_fouls',
       'away_corners', 'away_yellowCards', 'away_redCards',
       'away_total_assists', 'away_total_xAssists', 'away_total_key_passes',
       'away_total_xGoalsChain', 'away_total_xGoalsBuildup',
       'away_total_yellow_cards', 'away_total_red_cards',
       'away_total_blocked_shots', 'away_total_saved_shots', 'gameresult'],
      dtype='object')
```

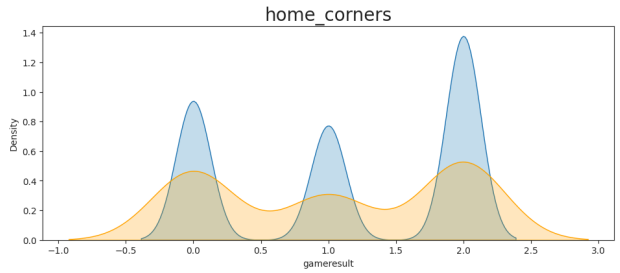
Difference in the distribution of a variable when another variable is with or without MV:

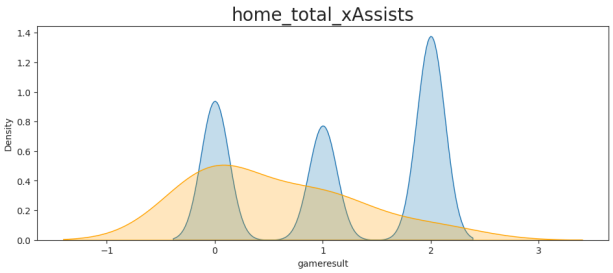

home_xGoals
Density
gameresult

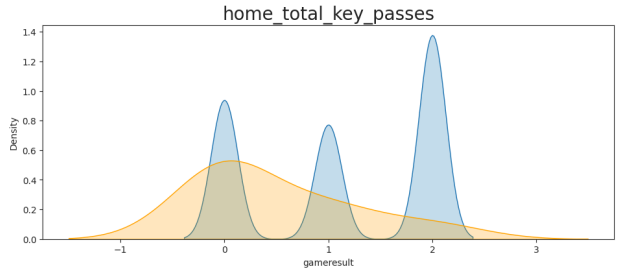

home_shots
Density
gameresult

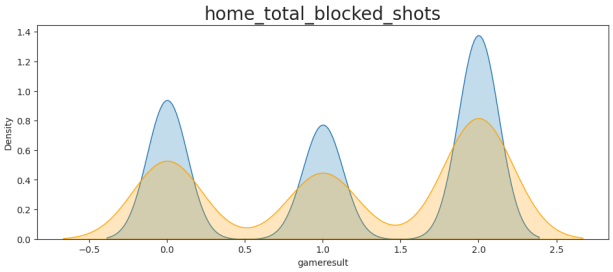

home_deep
Density
gameresult

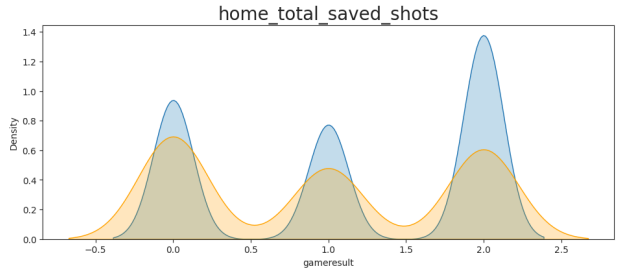

home_fouls
Density
gameresult

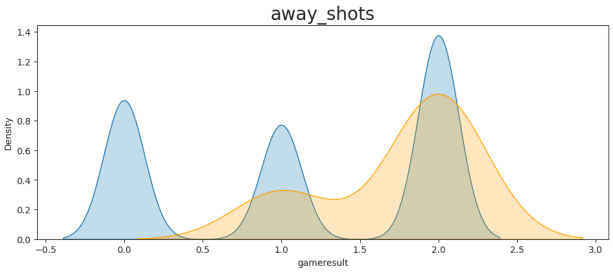

home_corners
Density
gameresult

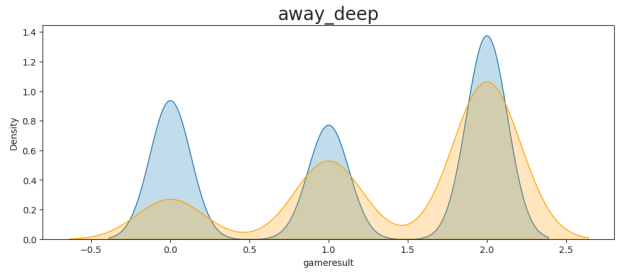

home_total_xAssists
Density
gameresult

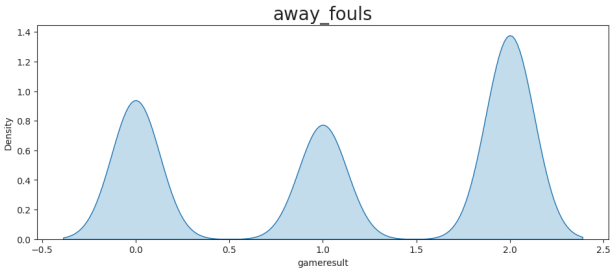

home_total_key_passes
Density
gameresult

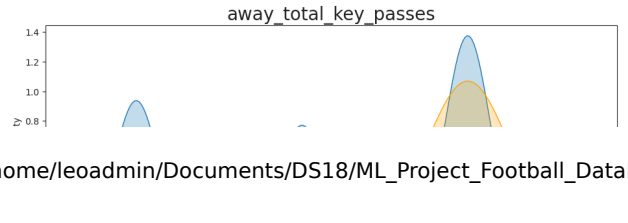

home_total_blocked_shots
Density
gameresult

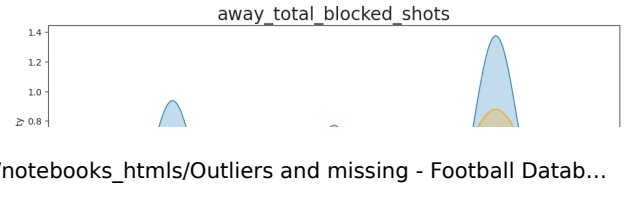

home_total_saved_shots
Density
gameresult

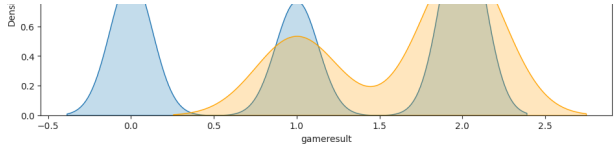

away_shots
Density
gameresult

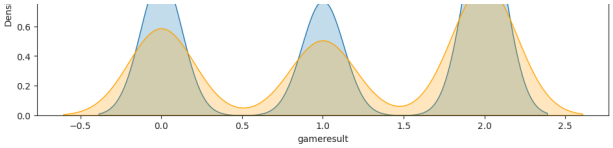

away_deep
Density
gameresult

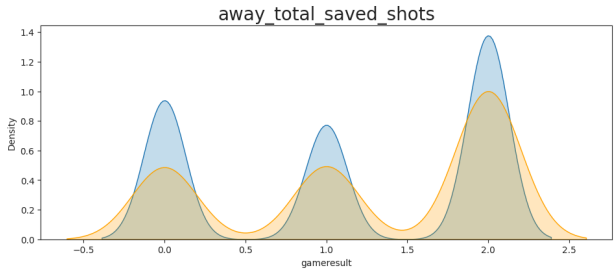

away_fouls
Density
gameresult

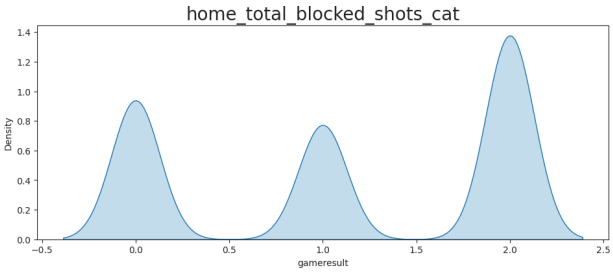

away_total_key_passes

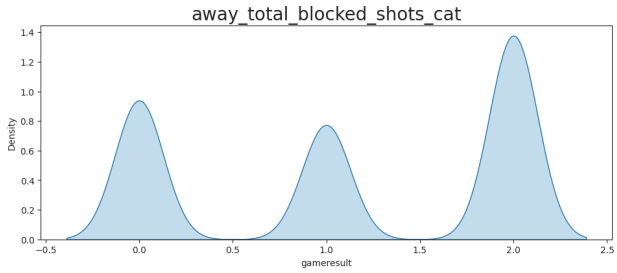

away_total_blocked_shots

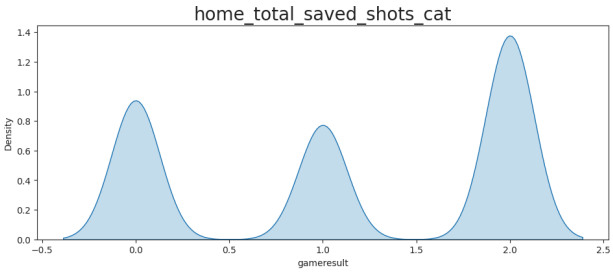


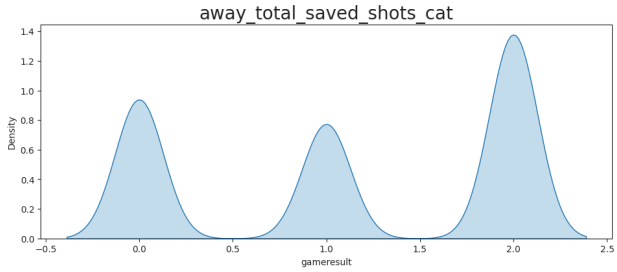










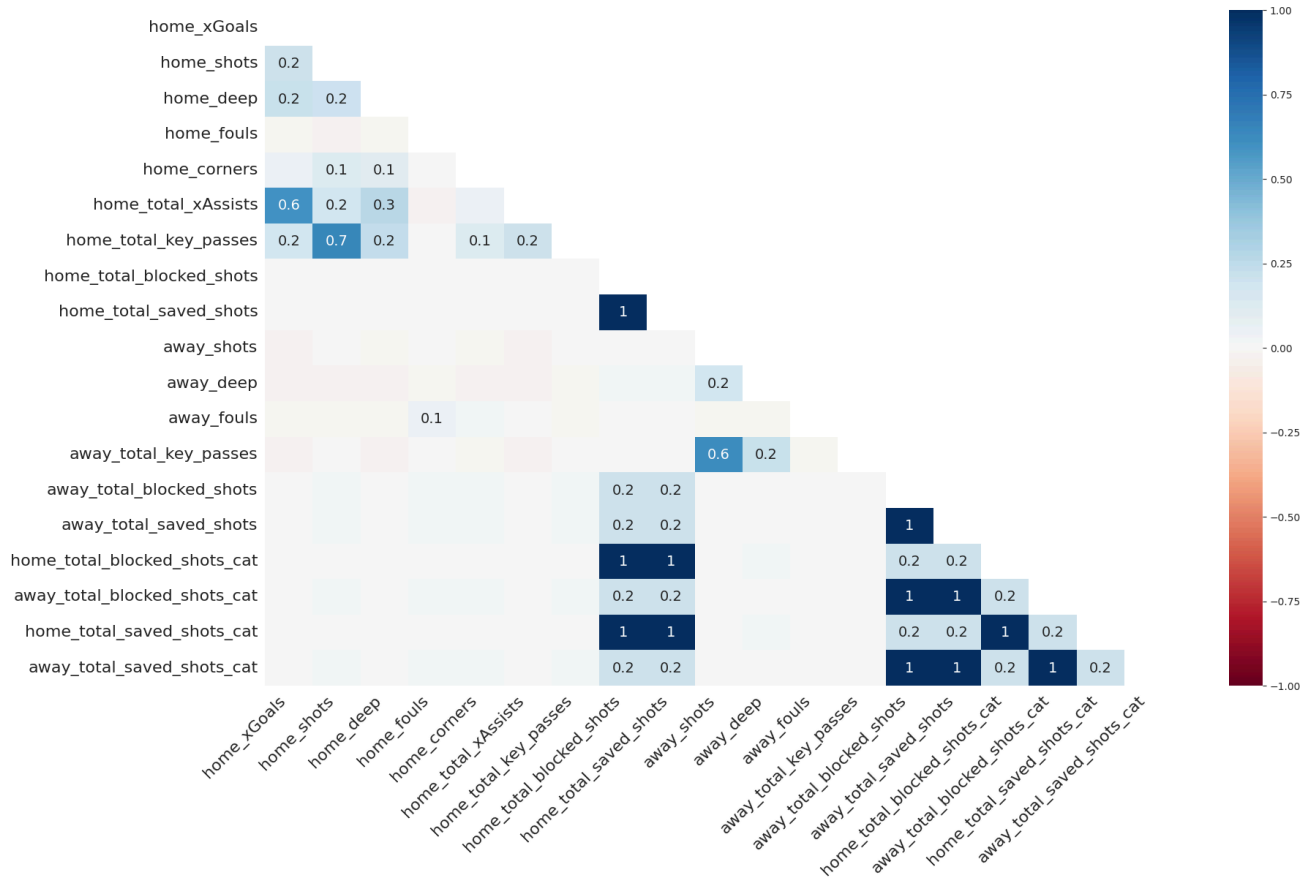


Exploring data the segnificunce of distribution change:

	Var	MV_pct	distribution_changed
0	gameID	0	-
1	leagueID	0	-
2	season	0	-
3	homeTeamID	0	-
4	awayTeamID	0	-
5	home_Goals	0	-
6	away_Goals	0	-
7	home_GoalsHalfTime	0	-
8	away_GoalsHalfTime	0	-
9	home_xGoals	276	+
10	home_shots	167	+
11	home_shotsOnTarget	0	-
12	home_deep	227	+
13	home_ppda	0	-
14	home_fouls	241	+
15	home_corners	143	+
16	home_yellowCards	0	-
17	home_redCards	0	-
18	home_total_assists	0	-
19	home_total_xAssists	346	+
20	home_total_key_passes	123	-
21	home_total_xGoalsChain	0	-
22	home_total_xGoalsBuildup	0	-
23	home_total_yellow_cards	0	-
24	home_total_red_cards	0	-
25	home_total_blocked_shots	3	-
26	home_total_saved_shots	3	-
27	away_xGoals	0	-
28	away_shots	161	+
29	away_shotsOnTarget	0	-
30	away_deep	423	+
31	away_ppda	0	-
32	away_fouls	81	-
33	away_corners	0	-
34	away_yellowCards	0	-
35	away_redCards	0	-

	Var	MV_pct	distribution_changed
36	away_total_assists	0	-
37	away_total_xAssists	0	-
38	away_total_key_passes	169	+
39	away_total_xGoalsChain	0	-
40	away_total_xGoalsBuildup	0	-
41	away_total_yellow_cards	0	-
42	away_total_red_cards	0	-
43	away_total_blocked_shots	8	-
44	away_total_saved_shots	8	-
45	gameresult	0	-

<Axes: >



	Var	MV_pct	distribution_changed	drop	MV_type
30	away_deep	423	+	No	MNAR
19	home_total_xAssists	346	+	No	MNAR
9	home_xGoals	276	+	No	MNAR
14	home_fouls	241	+	No	MNAR
12	home_deep	227	+	No	MNAR
38	away_total_key_passes	169	+	No	MNAR
10	home_shots	167	+	No	MNAR
28	away_shots	161	+	No	MNAR
15	home_corners	143	+	No	MNAR
20	home_total_key_passes	123	-	Yes	MCAR/MAR
32	away_fouls	81	-	Yes	MCAR/MAR
43	away_total_blocked_shots	8	-	Yes	MCAR/MAR
44	away_total_saved_shots	8	-	Yes	MCAR/MAR
25	home_total_blocked_shots	3	-	Yes	MCAR/MAR
26	home_total_saved_shots	3	-	Yes	MCAR/MAR

1. MCAR (Missing Completely At Random)

- Definition: The probability of a value being missing does not depend on the data itself or any other observed variables.
- Implication: You can safely use simple methods (e.g., dropping rows, mean/median imputation, KNN, MICE, etc.) without introducing strong biases.
- In Your Table: home_total_saved_shots might be MCAR or MAR (since it's labeled "MCAR/MAR"). If you trust it's MCAR, a straightforward imputation (like median or mode) or even dropping those rows might be acceptable—but your table says "drop = Yes," suggesting you decided to drop this variable altogether.

2. MAR (Missing At Random)

- Definition: The probability of missing data may depend on other observed variables, but not on the value of the variable itself.
- Implication: You can still do imputation, but you might want to incorporate other columns. For instance, multiple imputation (MICE) or a regression-based approach that uses other features to predict the missing values.
- In Your Table: home_total_saved_shots is also labeled "MCAR/MAR." If you suspect MAR, a more advanced approach (like MICE) can be used if you're not dropping the variable.

3. MNAR (Missing Not At Random)

- Definition: The missingness depends on the value of the variable itself. For example, if teams with high corners/fouls are more likely to omit that data.
- Implication: Standard imputation methods can be biased. You may need:
    1. Domain knowledge: Understand why it's missing.
    2. Indicator variable: Sometimes you create an extra column, e.g. col_missing = {0,1}, to capture the fact that a value was missing.
    3. Special modeling: If the data is systematically missing for certain outcomes, you might treat "missingness" as a separate category or use specialized MNAR models.
- In Your Table: The columns labeled "MNAR" (home_fouls, away_fouls, home_corners, away_corners, etc.) are trickier. If you suspect the missingness is related to the actual values (e.g., maybe referees are less likely to record certain stats when the number is extremely high or for certain matches), you cannot simply do mean/median imputation without risking bias.
    - One approach is to add an indicator for missing vs. non-missing, then for the non-missing subset, do an advanced approach or even treat them as unknown.
    - If you have no domain-based approach, you might end up ignoring or dropping those columns, but that can lose important info.

4. "distribution_changed" = "+"

- If distribution_changed is "+", it indicates the distribution of that feature changes significantly when missing values are removed or considered. This often means the missing data is systematically different from the observed data. Coupled with MNAR, it strongly suggests caution if you try to fill them with a naive approach.
- If it's "-," it means the distribution does not change significantly, so simpler approaches might be less biased.

5. "drop = Yes"

- This typically indicates you decided to drop that feature from the analysis entirely, either because it's too problematic or you have too few data points.
- In your table, home_total_saved_shots is marked "Yes" to drop. Possibly you decided the missingness is too frequent or unpredictable. Alternatively, you might have discovered it's not crucial to your model.

6. Practical Steps

1. Columns Marked MNAR:

    - Evaluate whether the missingness might be systematic. Possibly create an indicator column (col_is_missing) for them. Then for non-missing rows, you could do domain-based or advanced imputation. Or treat them as “missing = new category” if it’s a categorical scenario.

    - If the missingness is large and you have no good method, consider dropping them or acknowledging a potential bias.

2. Columns Marked MCAR/MAR:

    - Use typical imputation (mean/median/mode, KNN, MICE).

    - If it’s truly MCAR and the missing ratio is small, dropping rows might be okay. If it’s MAR, try a more advanced approach that uses other features to guess the missing values.

3. Distribution Changed:

    - If “+,” the presence of missing data is significantly altering the distribution. That’s a red flag for naive imputation. Possibly add a missing indicator, or investigate the pattern of missingness further.

4. heck Percentage:

    - Look at MV_pct (missing percentage). If it’s extremely high (like 50%+), you might not salvage that column. If it’s moderate, an advanced approach might help.

5. Domain Knowledge:

    - For example, if “corners” are missing only for certain tournaments or for certain teams, you might glean the reason behind the missingness. That helps you decide how to handle them.

6. Final Decision:

    - There’s no one-size-fits-all solution. Evaluate each column’s importance, the extent of missingness, and the mechanism (MNAR, MAR, MCAR). Then choose:

        - Drop the column (if not critical or if data is too incomplete).

        - Impute with advanced methods.

        - Create a missing indicator (particularly for MNAR).

        - Leave as is (missing can be an explicit category or value if that makes sense in your domain).

Summary

- MNAR: Hardest to handle. Often create a missing indicator or do specialized modeling.
- MAR: Imputation using other columns (e.g., MICE) is typically acceptable.
- MCAR: Simple approaches (drop or basic imputation) are relatively safe.
- distribution_changed = “+” means be extra cautious with naive imputation.
- If “drop=Yes,” you plan to remove that feature from further analysis.

By carefully combining these strategies—especially for MNAR columns—you can reduce bias and handle your missing data in a principled way.

Following the above, the idea of nulling the outliers in that dataset may cause a bias when using methods to fill the missing values. Therefore, only the real missing values will be dealt and another aproach will be applied on outliers - transform the features to categorical.

The best approach depends on the missingness mechanism and the impact on your analysis. Based on your tables:

1. For MNAR features (e.g., away_deep, home_total_xAssists, home_xGoals, home_fouls, home_deep, away_total_key_passes, home_shots, away_shots, home_corners) where the distribution changes when missing values are dropped, a simple imputation (like median) may introduce bias because the missingness is related to the value itself. For these you could:
    - Use a more sophisticated imputation method (e.g., multiple imputation by chained equations [MICE] or regression-based imputation) that uses other predictors in your data.
    - If the missing percentage is very low (1–3%), you might accept a median imputation as a practical solution—but keep in mind the potential bias.
2. For MCAR/MAR features that are marked “drop=Yes” (e.g., home_total_key_passes, away_fouls, away_total_blocked_shots, away_total_saved_shots, home_total_blocked_shots, home_total_saved_shots) with extremely low missing percentages (0–0.1%), you could:
    - Drop these features if they aren’t critical.
    - Or, if you prefer to keep them, impute using the median since their missingness is likely random and their proportions are very small.

In summary, if you need to keep the MNAR features, using an advanced imputation method

(MICE or model-based imputation) is preferred. For the MCAR/MAR features, a simple median imputation (or dropping them altogether) should be sufficient due to their very low missing rate.

Data Imputation

	home_xGoals	home_shots	home_deep	home_fouls	home_corners	home_total_xAssists
0	0.627539	9.0	4.0	12.0	1.0	0.284979
1	0.876106	11.0	11.0	13.0	6.0	0.419975
2	0.604226	10.0	5.0	7.0	8.0	0.549139
3	2.568030	19.0	5.0	13.0	6.0	1.727543
4	1.130760	17.0	5.0	14.0	1.0	0.416638
...	...	...	...	...	...	...
12675	1.411190	15.0	17.0	8.0	9.0	0.971853
12676	1.198190	10.0	3.0	11.0	5.0	0.855524
12677	1.332690	12.0	10.0	11.0	4.0	1.151649
12678	1.460500	19.0	6.0	13.0	9.0	1.265829
12679	0.323960	6.0	1.0	17.0	2.0	0.074636

12680 rows × 19 columns

```
Index(['home_xGoals', 'home_shots', 'home_deep', 'home_fouls', 'home_corners',
       'home_total_xAssists', 'home_total_key_passes',
       'home_total_blocked_shots', 'home_total_saved_shots', 'away_shots',
       'away_deep', 'away_fouls', 'away_total_key_passes',
       'away_total_blocked_shots', 'away_total_saved_shots',
       'home_total_blocked_shots_cat', 'away_total_blocked_shots_cat',
       'home_total_saved_shots_cat', 'away_total_saved_shots_cat'],
      dtype='object')
```

	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
0	81	1	2015	2015-08-08 15:45:00	89	82	1	0
1	82	1	2015	2015-08-08 18:00:00	73	71	0	1
2	83	1	2015	2015-08-08 18:00:00	72	90	2	2
3	84	1	2015	2015-08-08 18:00:00	75	77	4	2
4	85	1	2015	2015-08-08 18:00:00	79	78	1	3
...	...	...	...	...	...	...	...	...
12675	16131	5	2020	2021-05-23 19:00:00	168	166	1	2
12676	16132	5	2020	2021-05-23 19:00:00	177	176	1	2
12677	16133	5	2020	2021-05-23 19:00:00	163	235	2	0
12678	16134	5	2020	2021-05-23 19:00:00	175	181	0	1
12679	16135	5	2020	2021-05-23 19:00:00	225	179	1	1

12680 rows × 63 columns

```
gameID                          0
leagueID                        0
season                          0
date                            0
homeTeamID                      0
                               ..
away_Goals_cat                  0
home_total_blocked_shots_cat    3
away_total_blocked_shots_cat    8
home_total_saved_shots_cat      3
away_total_saved_shots_cat      8
Length: 63, dtype: int64
```

	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
0	81	1	2015	2015-08-08 15:45:00	89	82	1	0
1	82	1	2015	2015-08-08 18:00:00	73	71	0	1
2	83	1	2015	2015-08-08 18:00:00	72	90	2	2
3	84	1	2015	2015-08-08 18:00:00	75	77	4	2
4	85	1	2015	2015-08-08 18:00:00	79	78	1	3
...	...	...	...	...	...	...	...	...
12675	16131	5	2020	2021-05-23 19:00:00	168	166	1	2
12676	16132	5	2020	2021-05-23 19:00:00	177	176	1	2
12677	16133	5	2020	2021-05-23 19:00:00	163	235	2	0
12678	16134	5	2020	2021-05-23 19:00:00	175	181	0	1
12679	16135	5	2020	2021-05-23 19:00:00	225	179	1	1

12680 rows × 63 columns

```
gameID                          0
leagueID                        0
season                          0
date                            0
homeTeamID                      0
                               ..
away_Goals_cat                  0
home_total_blocked_shots_cat    3
away_total_blocked_shots_cat    8
home_total_saved_shots_cat      3
away_total_saved_shots_cat      8
Length: 63, dtype: int64
```

need to deal with the missing values for the categories features using the fiiling of the corespond feature

	gameID	leagueID	season	date	homeTeamID	awayTeamID	home_Goals	away_Goals
0	81	1	2015	2015-08-08 15:45:00	89	82	1	0
1	82	1	2015	2015-08-08 18:00:00	73	71	0	1
2	83	1	2015	2015-08-08 18:00:00	72	90	2	2
3	84	1	2015	2015-08-08 18:00:00	75	77	4	2
4	85	1	2015	2015-08-08 18:00:00	79	78	1	3
...	...	...	...	...	...	...	...	...
12675	16131	5	2020	2021-05-23 19:00:00	168	166	1	2
12676	16132	5	2020	2021-05-23 19:00:00	177	176	1	2
12677	16133	5	2020	2021-05-23 19:00:00	163	235	2	0
12678	16134	5	2020	2021-05-23 19:00:00	175	181	0	1
12679	16135	5	2020	2021-05-23 19:00:00	225	179	1	1

12680 rows × 63 columns

```
gameID                          0
leagueID                        0
season                          0
date                            0
homeTeamID                      0
                               ..
away_Goals_cat                  0
home_total_blocked_shots_cat    0
away_total_blocked_shots_cat    0
home_total_saved_shots_cat      0
away_total_saved_shots_cat      0
Length: 63, dtype: int64
```